

Fast Computation of Zeros of Polynomial Systems with Bounded Degree under Finite-precision

<p>Irénée Briquel Dept. of Mathematics City University of Hong Kong HONG KONG e-mail: irenee.briquel@gmail.com</p>	<p>Felipe Cucker Dept. of Mathematics City University of Hong Kong HONG KONG e-mail: macucker@cityu.edu.hk</p>
<p>Javier Peña Carnegie Mellon University Tepper School of Business PA, USA e-mail: jfp@andrew.cmu.edu</p>	<p>Vera Roshchina Centro de Matemática e Aplicações Universidade de Évora PORTUGAL e-mail: vera.roshchina@gmail.com</p>

Abstract. A solution for Smale’s 17th problem, for the case of systems with bounded degree was recently given. This solution, an algorithm computing approximate zeros of complex polynomial systems in average polynomial time, assumed infinite precision. In this paper we describe a finite-precision version of this algorithm. Our main result shows that this version works within the same time bounds and requires a precision which, on the average, amounts to a polynomial amount of bits in the mantissa of the intervening floating-point numbers.

Keywords: Smale’s 17th problem, finite-precision, polynomial systems.

AMS Subject Classification: 65G50, 65H10, 65Y20.

1 Introduction

The 17th of the problems for the 21st century posed by Steve Smale [19] asks for an algorithm computing an approximate zero of a polynomial system in average polynomial time.

The problem had occupied Smale during the 1990’s and led him, together with Mike Shub, to a series of papers [13, 14, 15, 17, 16, 12] —known as *the Bézout series*— where a number of ideas and results approaching a solution for the 17th problem were proposed. These ideas are at the core of all further research done on Smale’s problem.

Paramount within this research is the work of Carlos Beltrán and Luis Miguel Pardo [2, 3, 4] who provided a randomized algorithm computing the desired approximate zero in average expected polynomial time. Here the word “average” refers to expectation over the input data and the word “expected” to expectation over the random choices made by the

algorithm¹. One can say that they gave a probabilistic solution to Smale’s 17th problem. Further results, including a deterministic algorithm working in average time $N^{\mathcal{O}(\log \log N)}$ —referred to as “nearly polynomial” — are given in [6]. This deterministic algorithm, when restricted to systems with bounded (or even moderately growing) degree, becomes an average polynomial-time algorithm, referred to in [6] as MD.

All the mentioned work (as well as all the work on Smale’s 17th problem not mentioned above) assume infinite precision. As Beltrán and Pardo put it in [4, p. 6]

With the assumption of exact arithmetic [...] the homotopy method [...] is guaranteed to produce an approximate zero of f .

This statement begs the question, what can one do if (as it happens with digital computers) only finite precision is available²? The goal of the present paper is to give an answer for systems of moderate degree.

The distinctive feature of a finite precision algorithm is the presence of a real number $u \in (0, 1)$, called *round-off unit*, with the property that all occurring numbers x in the computation are replaced by a number $r_u(x)$ (the rounding of x) such that $|x - r_u(x)| \leq u|x|$ (a more detailed account on finite-precision computations is in §3.1). Algorithms where u remains fixed through the computation are said to have *fixed precision*. Otherwise, they have *variable precision*. In this paper, we describe and analyze finite-precision versions for both settings —we denote them by MDfix and MDvar, respectively— of algorithm MD. The rationale for the consideration of both settings will be made clear soon enough.

Our results are of a probabilistic nature (often referred to as an “average-case analysis”). They therefore require a probability measure on the space of data. We next describe the measures we use.

Let $\mathcal{H}_{(\mathbf{d})}$ denote the linear space of complex polynomial systems $f = (f_1, \dots, f_n)$ with f_i homogeneous of degree d_i in $n + 1$ variables. Given a round-off unit \bar{u} , each system $f \in \mathcal{H}_{(\mathbf{d})}$ is rounded to a system $r_{\bar{u}}(f)$. The data space corresponding to this precision is therefore

$$\mathcal{H}_{\bar{u}} := \{r_{\bar{u}}(f) \mid f \in \mathcal{H}_{(\mathbf{d})}\}.$$

Clearly we have a surjective function $r_{\bar{u}} : \mathcal{H}_{(\mathbf{d})} \rightarrow \mathcal{H}_{\bar{u}}$. The measure we endow $\mathcal{H}_{\bar{u}}$ with is the push-out $\nu_{\bar{u}}$ of the standard Gaussian μ in $\mathcal{H}_{(\mathbf{d})}$. That is, we endow $\mathcal{H}_{(\mathbf{d})}$ with a standard Gaussian distribution (with respect to the Bombieri-Weyl basis in $\mathcal{H}_{(\mathbf{d})}$, see §2.1 for details) and we define, for all Borelian subset $A \subseteq \mathcal{H}_{\bar{u}}$, $\nu_{\bar{u}}(A) := \mu(r_{\bar{u}}^{-1}(A))$. With the measure $\nu_{\bar{u}}$ at hand we can state our first main result.

Theorem A. *There exists a fixed precision algorithm MDfix satisfying the following. For a random input $f \in \mathcal{H}_{\bar{u}}$, algorithm MDfix returns an approximate zero of f with probability at least*

$$1 - \mathcal{O}\left(\frac{D^3 N(n+1)^{D+1}}{\log(1/\bar{u})}\right).$$

¹Although technically similar, there is a remarkable difference between the probability distributions considered. The one for the input data is, explicitly or otherwise, claiming some closeness to the distribution of data “in practice.” The only requirement for the distribution for the random choices of the algorithm is, in contrast, that it will be efficiently computable. An undisputed merit of the work of Beltrán and Pardo is to come up with one distribution which is so and, at the same time, allows one to derive complexity bounds.

²Incidentally, finite precision analysis for algorithms dealing with multivariate polynomial systems was pioneered by Steve Smale, dragging on the way one of the authors of the present paper, in [8].

Otherwise, `MDFix` returns a failure message. The number of arithmetic operations performed is bounded as

$$\mathcal{O}\left(\frac{1}{n\sqrt{D}(\log N + D + n^2)\bar{u}}\right).$$

Here $N = \dim_{\mathbb{C}} \mathcal{H}_{(\mathbf{d})}$ denotes the size of the input systems and $D := \max\{d_1, \dots, d_n\}$.

While the consideration of fixed precision is a realistic approach, a result such as Theorem A is not without shortcomings. Note that `MDFix` does not always return an approximate zero of its input f , and it fails to do so with positive probability. And, unfortunately, the lower bound for the probability of success shown in Theorem A becomes meaningless when N grows. In other words, a fixed precision \bar{u} puts limits on the size of the systems for which one can expect `MDFix` to succeed.

The relationship between input size, round-off unit, and probability of success implicit in Theorem A can be better expressed with the use of variable precision and the possibility of continuously adjusting the input reading to the current round-off unit. This is a less realistic setting but it pays off in terms of understanding. We next see how.

We assume variable precision. That is, algorithms now start with an initial round-off unit and they have the capability to refine this parameter as the execution proceeds. In this context we will be interested in the smallest value u_* attained by the round-off unit during the execution. A bound on u_* amounts to a bound on the number of bits (or digits) required to store floating point approximations of the complex numbers occurring during the computation and, in this sense, is related to the bit-cost of performing the computation. In fact, the maximum number of bits we will need for each such approximation is essentially $\lceil \log u_* \rceil$.

An issue naturally raised by the assumption of variable precision is the space of data from which algorithms will take their inputs. The spaces $\mathcal{H}_{\bar{u}}$ are appropriate for the fixed precision setting: systems f are read with the precision used throughout the computation. They do not appear to be so for the variable precision context. But finite-precision algorithms cannot take inputs with infinite-precision entries. An elegant solution for this situation is the consideration of black-boxes. These are theoretical devices which, as we said, are not realistic. But they do allow for the statement of results highlighting in a clear manner the relationship between precision needed and input size.

In what follows, to every $f \in \mathcal{H}_{(\mathbf{d})}$ we associate a routine `read_inputf` such that `read_inputf()` returns an approximation of f with the current round-off unit u . It is this routine what is given as input to our variable precision algorithm `MDVar`. Because of the bijection between this set of routines and the space $\mathcal{H}_{(\mathbf{d})}$ we may (and will) abuse language and take $\mathcal{H}_{(\mathbf{d})}$ as the space of data for `MDVar`. In particular, we will endow this space of data with the standard Gaussian mentioned above, which we will denote by $N(0, \text{Id})$.

Our second main result is the following.

Theorem B. *There exists a variable precision algorithm `MDVar` satisfying the following. When f is randomly chosen from $N(0, \text{Id})$, algorithm `MDVar` on input f stops almost surely, and when it does so, returns an approximate zero of f . The number of arithmetic operations $\text{cost}_{\text{MDVar}}(f)$ of `MDVar` on input f is bounded on the average as*

$$\mathbb{E}_{f \sim N(0, \text{Id})} \text{cost}_{\text{MDVar}}(f) = \mathcal{O}(D^3 N^2 (n+1)^{D+1}).$$

Furthermore, the finest precision $u_*(f)$ used by MDVar on input f is bounded on the average as

$$\mathbb{E}_{f \sim N(0, \text{Id})} \log |u_*(f)| = \mathcal{O}(D^3 N(n+1)^{D+1})$$

and as a consequence, when D is bounded, the bit-cost of MDVar is, on the average, polynomial in the size N of the input.

Before proceeding with the technicalities, a couple of remarks are in place.

(1) As mentioned above, we confirm that Theorem B adds understanding to the situation depicted in Theorem A. Indeed, in the variable precision context the algorithm returns an approximate zero almost surely (that is, the probability of failure is now zero). Furthermore, the complexity (understood as number of arithmetic operations performed) of MDVar remains essentially the same as that of MD. Finally, the relationship between precision and input size is made clear: we exhibit polynomial bounds (in the input size N) for the expected (over random input systems f) number of bits necessary to carry out the computation.

(2) Both Theorems A and B show only the existence of algorithms MDFix and MDVar, respectively. We do not fully exhibit these algorithms in this paper. For such a complete description we should provide the exact values of some constants occurring within the ‘big Oh’ notation in a few intermediate results in our development. This is certainly possible but we believe that doing so would only degrade our exposition. Because, on the one hand, it would bring undue focus on a marginal issue and, on the other hand, the length of the exposition would unavoidably increase in a non-trivial way. We follow in this sense a well established tradition in finite-precision analyses, by which the goal is the understanding of the magnitude of the precision needed by an algorithm. In particular, we do not intend algorithms MDFix and MDVar to be actually implemented. They are simply vehicles to understand the behavior of the (already implemented) algorithm MD and in this sense we make ours the words of Wilkinson quoted by Higham to open the tenth chapter of [10]:

All too often, too much attention is paid to the precise error bound that has been established. The main purpose of such an analysis is either to establish the essential numerical stability of an algorithm or to show why it is unstable and in doing so to expose what sort of change is necessary to make it stable. The precise error bound is not of great importance.

2 Preliminaries

2.1 Setting and Notation

For $d \in \mathbb{N}$ we denote by H_d the subspace of $\mathbb{C}[X_0, \dots, X_n]$ of homogeneous polynomials of degree d . For $f \in H_d$ we write

$$f(X) = \sum_{\alpha} \binom{d}{\alpha}^{1/2} a_{\alpha} X^{\alpha}$$

where $\alpha = (\alpha_0, \dots, \alpha_n)$ is assumed to range over all multi-indices such that $|\alpha| = \sum_{k=0}^n \alpha_k = d$, $\binom{d}{\alpha}$ denotes the multinomial coefficient, and $X^{\alpha} := X_0^{\alpha_0} X_1^{\alpha_1} \dots X_n^{\alpha_n}$. That is, we take for basis of the linear space H_d the *Bombieri-Weyl* basis consisting of the monomials $\binom{d}{\alpha}^{1/2} X^{\alpha}$. A reason to do so is that the Hermitian inner product associated to this basis is unitarily

invariant. That is, if $g \in H_d$ is given by $g(x) = \sum_{\alpha} \binom{d}{\alpha}^{1/2} b_{\alpha} X^{\alpha}$, then the canonical Hermitian inner product

$$\langle f, g \rangle = \sum_{|\alpha|=d} a_{\alpha} \overline{b_{\alpha}}$$

satisfies, for all elements ν in the unitary group $\mathcal{U}(n+1)$, that

$$\langle f, g \rangle = \langle f \circ \nu, g \circ \nu \rangle.$$

Fix $d_1, \dots, d_n \in \mathbb{N} \setminus \{0\}$ and let $\mathcal{H}_{(\mathbf{d})} = H_{d_1} \times \dots \times H_{d_n}$ be the vector space of polynomial systems $f = (f_1, \dots, f_n)$ with $f_i \in \mathbb{C}[X_0, \dots, X_n]$ homogeneous of degree d_i . The space $\mathcal{H}_{(\mathbf{d})}$ is naturally endowed with a Hermitian inner product $\langle f, g \rangle = \sum_{i=1}^n \langle f_i, g_i \rangle$. We denote by $\|f\|$ the corresponding norm of $f \in \mathcal{H}_{(\mathbf{d})}$.

We let $N := \dim_{\mathbb{C}} \mathcal{H}_{(\mathbf{d})}$, $D := \max_i d_i$, and $\mathcal{D} := \prod_i d_i$. Also, in the rest of this paper, we assume $d_i \geq 2$ for all $i \leq n$ (linear equations can be easily eliminated). In particular, $D \geq 2$.

Let $\mathbb{P}^n := \mathbb{P}(\mathbb{C}^{n+1})$ denote the complex projective space associated to \mathbb{C}^{n+1} and $S(\mathcal{H}_{(\mathbf{d})})$ the unit sphere of $\mathcal{H}_{(\mathbf{d})}$. These are smooth manifolds that naturally carry the structure of a Riemannian manifold (for \mathbb{P}^n the metric is called Fubini-Study metric). We will denote by $d_{\mathbb{P}}$ and $d_{\mathbb{S}}$ their Riemannian distances which, in both cases, amount to the angle between the arguments. Specifically, for $x, y \in \mathbb{P}^n$ one has

$$\cos d_{\mathbb{P}}(x, y) = \frac{|\langle x, y \rangle|}{\|x\| \|y\|}. \quad (1)$$

Occasionally, for $f, g \in \mathcal{H}_{(\mathbf{d})} \setminus \{0\}$, we will abuse language and write $d_{\mathbb{S}}(f, g)$ to denote this angle, that is, the distance $d_{\mathbb{S}}\left(\frac{f}{\|f\|}, \frac{g}{\|g\|}\right) = d_{\mathbb{S}}(f, g)$. We define the *solution variety* to be

$$V_{\mathbb{P}} := \{(f, \zeta) \in \mathcal{H}_{(\mathbf{d})} \times \mathbb{P}^n \mid f \neq 0 \text{ and } f(\zeta) = 0\}.$$

This is a smooth submanifold of $\mathcal{H}_{(\mathbf{d})} \times \mathbb{P}^n$ and hence also carries a Riemannian structure. We denote by $V_{\mathbb{P}}(f)$ the zero set of $f \in \mathcal{H}_{(\mathbf{d})}$ in \mathbb{P}^n .

By Bézout's Theorem, $V_{\mathbb{P}}(f)$ contains \mathcal{D} points for almost all f . Let $Df(\zeta)|_{T_{\zeta}}$ denote the restriction of the derivative of $f: \mathbb{C}^{n+1} \rightarrow \mathbb{C}^n$ at ζ to the tangent space $T_{\zeta} := \{v \in \mathbb{C}^{n+1} \mid \langle v, \zeta \rangle = 0\}$ of \mathbb{P}^n at ζ . The *subvariety of ill-posed pairs* is defined as

$$\Sigma'_{\mathbb{P}} := \{(f, \zeta) \in V_{\mathbb{P}} \mid \text{rank } Df(\zeta)|_{T_{\zeta}} < n\}.$$

Note that $(f, \zeta) \notin \Sigma'_{\mathbb{P}}$ means that ζ is a simple zero of f . In this case, by the implicit function theorem, the projection $V_{\mathbb{P}} \rightarrow \mathcal{H}_{(\mathbf{d})}, (g, x) \mapsto g$ can be locally inverted around (f, ζ) . The image Σ of $\Sigma'_{\mathbb{P}}$ under the projection $V_{\mathbb{P}} \rightarrow \mathcal{H}_{(\mathbf{d})}$ is called the *discriminant variety*.

2.2 Approximate Zeros, Complexity and Data Distribution

In [11], Mike Shub introduced the following projective version of Newton's method. We associate to $f \in \mathcal{H}_{(\mathbf{d})}$ (with $Df(x)$ of rank n for some x) a map $N_f: \mathbb{C}^{n+1} \setminus \{0\} \rightarrow \mathbb{C}^{n+1} \setminus \{0\}$ defined (almost everywhere) by

$$N_f(x) = x - Df(x)|_{T_x}^{-1} f(x).$$

Note that $N_f(x)$ is homogeneous of degree 0 in f and of degree 1 in x so that N_f induces a rational map from \mathbb{P}^n to \mathbb{P}^n (which we will still denote by N_f) and this map is invariant under multiplication of f by constants.

We note that $N_f(x)$ can be computed from f and x very efficiently: since the Jacobian $Df(x)$ can be evaluated with $\mathcal{O}(N)$ arithmetic operations [1], one can do with a total of $\mathcal{O}(N + n^3) = \mathcal{O}(N)$ arithmetic operations, the equality since $d_i \geq 2$ implies $N = \Omega(n^3)$.

It is well-known that when x is sufficiently close to a simple zero ζ of f , the sequence of Newton iterates beginning at x will converge quadratically fast to ζ . This property led Steve Smale to define the following intrinsic notion of approximate zero.

Definition 2.1. By an *approximate zero* of $f \in \mathcal{H}_{(\mathbf{d})}$ associated with a zero $\zeta \in \mathbb{P}^n$ of f we understand a point $x \in \mathbb{P}^n$ such that the sequence of Newton iterates (adapted to projective space)

$$x_{i+1} := N_f(x_i)$$

with initial point $x_0 := x$ converges immediately quadratically to ζ , i.e.,

$$d_{\mathbb{P}}(x_i, \zeta) \leq \left(\frac{1}{2}\right)^{2^i - 1} d_{\mathbb{P}}(x_0, \zeta)$$

for all $i \in \mathbb{N}$.

It is this notion of approximation that is referred to in the statement of Smale's 17th problem.

The last notion necessary to formally state Smale's problem is that of 'average cost'. For the cost of a computation Smale proposes the number of arithmetic operations (this includes comparisons and possibly square roots) performed during the computation. In the case of a finite-precision algorithm one needs to multiply this number by the largest number of bits (or digits) necessary to approximate the complex numbers occurring during the computation.

The word 'average' refers to the standard normal distribution for the data (input) system $f \in \mathcal{H}_{(\mathbf{d})}$. Recall, we express an element $f \in \mathcal{H}_{(\mathbf{d})}$ as a linear combination of the monomials in the Bombieri-Weyl basis. The standard normal distribution corresponds to choosing the coefficients in this combination independently and identically distributed from the centered Gaussian distribution on \mathbb{C} (which in turn amounts to draw real and imaginary parts independently from the centered Gaussian distribution on \mathbb{R}). We denote this distribution on $\mathcal{H}_{(\mathbf{d})}$ by $N(0, \text{Id})$.

Hence, if $\text{cost}(f)$ denotes the cost of computing an approximate zero for f with a given algorithm then the average cost of this algorithm, for inputs in $\mathcal{H}_{(\mathbf{d})}$, is given by the expected value

$$\mathbb{E}_{f \sim N(0, \text{Id})} \text{cost}(f).$$

We remark that if the cost is homogeneous of degree zero, that is, if $\text{cost}(f) = \text{cost}(\lambda f)$ for all $\lambda \neq 0$, then the expectation above is the same as the expectation with f drawn from the uniform distribution on the unit sphere $S(\mathcal{H}_{(\mathbf{d})})$.

Smale's 17th problem asks for an algorithm computing an approximate zero (in the sense of Definition 2.1) with average cost (for the cost and data distribution described above) bounded by $N^{\mathcal{O}(1)}$.

2.3 Condition Numbers

How close need x to be from ζ to be an approximate zero? This depends on how well conditioned the zero ζ is.

For $f \in \mathcal{H}_{(\mathbf{d})}$ and $x \in \mathbb{C}^{n+1} \setminus \{0\}$ we define the (*normalized*) *condition number* $\mu_{\text{norm}}(f, x)$ by

$$\mu_{\text{norm}}(f, x) := \|f\| \left\| (Df(x)|_{T_x})^{-1} \mathbf{diag} \left(\sqrt{d_1} \|x\|^{d_1-1}, \dots, \sqrt{d_n} \|x\|^{d_n-1} \right) \right\|,$$

where the right-hand side norm denotes the spectral norm and $\mathbf{diag}(a_i)$ denotes the diagonal matrix with entries a_i . Note that $\mu_{\text{norm}}(f, x)$ is homogeneous of degree 0 in both arguments, hence it is well defined for $(f, x) \in S(\mathcal{H}_{(\mathbf{d})}) \times \mathbb{P}^n$. Also, it is well known (see [5, Ch. 12, Corollary 3]) that $\mu_{\text{norm}}(f, x) \geq 1$.

The following result (essentially, a γ -Theorem in Smale's theory of estimates for Newton's method [18]) quantifies our claim above (see [6] for its proof).

Theorem 2.2. *Assume $f(\zeta) = 0$ and $d_{\mathbb{P}}(x, \zeta) \leq \frac{\nu_0}{D^{3/2} \mu_{\text{norm}}(f, \zeta)}$ where $\nu_0 := 3 - \sqrt{7} \approx 0.3542$. Then x is an approximate zero of f associated with ζ . \square*

The next result, Proposition 4.1 from [6], gives bounds on the variation of the condition number $\mu_{\text{norm}}(f, x)$ when f and x vary.

Proposition 2.3. *Assume $D \geq 2$. Let $0 < \varepsilon \leq 0.13$ be arbitrary and $C \leq \frac{\varepsilon}{5.2}$. For all $f, g \in S(\mathcal{H}_{(\mathbf{d})})$ and all $x, y \in \mathbb{C}^{n+1}$, if $d_{\mathbb{S}}(f, g) \leq \frac{C}{D^{1/2} \mu_{\text{norm}}(f, x)}$ and $d_{\mathbb{P}}(x, y) \leq \frac{C}{D^{3/2} \mu_{\text{norm}}(f, x)}$, then*

$$\frac{1}{1 + \varepsilon} \mu_{\text{norm}}(g, y) \leq \mu_{\text{norm}}(f, x) \leq (1 + \varepsilon) \mu_{\text{norm}}(g, y). \quad \square$$

In what follows, we will fix the constants $\varepsilon := 0.13$ and $C := \frac{\varepsilon}{5.2} = 0.025$.

We also introduce the *mean square condition number* of q given by

$$\mu_2^2(q) := \frac{1}{D} \sum_{\zeta: q(\zeta)=0} \mu_{\text{norm}}^2(q, \zeta). \quad (2)$$

2.4 An Adaptive Homotopy Continuation

Suppose that we are given an input system $f \in S(\mathcal{H}_{(\mathbf{d})})$ and a pair $(g, \zeta) \in V_{\mathbb{P}}$, where g is also in the unit sphere and such that f and g are \mathbb{R} -linearly independent. Let $\alpha = d_{\mathbb{S}}(f, g)$. Remark that one can compute α as

$$\alpha = 2 \arcsin \left(\frac{\|f - g\|}{2} \right). \quad (3)$$

Consider the line segment $E_{g,f}$ in $\mathcal{H}_{(\mathbf{d})}$ with endpoints g and f . We parameterize this segment by writing

$$E_{g,f} = \{q_\tau \in \mathcal{H}_{(\mathbf{d})} \mid \tau \in [0, 1]\}$$

with q_τ being the only point in $E_{g,f}$ such that $d_{\mathbb{S}}(g, q_\tau) = \tau\alpha$. Explicitly, as remarked in [6], we have $q_\tau = t f + (1 - t)g$, where $t = t(\tau)$ is given by

$$t(\tau) = \frac{1}{\sin \alpha \cot(\tau\alpha) - \cos \alpha + 1}. \quad (4)$$

If $E_{g,f} \cap \Sigma = \emptyset$, and hence almost surely, this segment can be lifted to a path given by a continuous function $[0, 1] \rightarrow V_{\mathbb{P}}$ mapping $\tau \mapsto (q_\tau, \zeta_\tau)$.

In order to find an approximation of the zero ζ_1 of $f = q_1$ we may start with the zero $\zeta = \zeta_0$ of $g = q_0$ and numerically follow the path (q_τ, ζ_τ) by subdividing $[0, 1]$ into points $0 = \tau_0 < \tau_1 < \dots < \tau_k = 1$ and by successively computing approximations x_i of ζ_{τ_i} by Newton's method.

This course of action is the one proposed in the Bézout series and further adopted in [2, 3, 4, 6]. The (infinite precision) continuation procedure is the following (here $\lambda = \frac{C(1-\varepsilon)}{2(1+\varepsilon)^4} \approx 6.67 \cdot 10^{-3}$, see [6]).

```

Algorithm ALH
input  $f, g, \zeta$ 
 $\#\# (g, \zeta) \in V, f \neq g \#\#$ 
 $\alpha := d_{\mathbb{S}}(f, g), \tau := 0, q_\tau := g$ 
repeat
   $\Delta\tau := \frac{\lambda}{\alpha D^{3/2} \mu_{\text{norm}}^2(q_\tau, x)}$ 
   $\tau := \min\{1, \tau + \Delta\tau\}$ 
   $q_\tau := t(\tau)f + (1 - t(\tau))g$ 
   $x := N_{\tilde{q}_\tau}(x)$ 
   $x := x/\|x\|$ 
until  $\tau = 1$ 
RETURN  $x$ 

```

Note that the step-length $\Delta\tau$ depends on $\mu_{\text{norm}}(q_\tau, x)$. Hence, the adaptiveness.

The algorithm MD (Moderate Degree) from [6] is a direct application of ALH having as initial pair (g, ζ) the pair (\overline{U}, z_1) , where $\overline{U} = (\overline{U}_1, \dots, \overline{U}_n) \in S(\mathcal{H}_{(\mathbf{d})})$ with $\overline{U}_i = \frac{1}{\sqrt{2n}}(X_0^{d_i} - X_i^{d_i})$ and $z_1 = \frac{1}{\sqrt{n+1}}(1, \dots, 1)$.

```

Algorithm MD
input  $f \in \mathcal{H}_{(\mathbf{d})}$ 
run ALH on input  $(f, \overline{U}, z_1)$ 

```

2.5 Roadmap

Theorems A and B are proved by designing finite-precision versions of algorithm ALH which take into account the errors due to the use of finite-precision. The variable precision version ALHVar is described in detail in Section 4. In particular, its main properties are shown in Theorem 4.3 in this section. Proposition 4.5 —a finite precision version of the inductive proof of [6, Theorem 3.1]— provides the backbone for the proof of Theorem 4.3.

Once with Theorem 4.3 at hand, the proof of Theorem B is carried out in a more or less straightforward manner in Section 5.

The use of fixed precision poses less demands in algorithmic design (the issue of round-off unit updating now becoming irrelevant). Algorithm MDFix is therefore a simplification of MDVar, which we describe in Section 6 together with the proof of Theorem A.

As just mentioned, the backbone of all this development is Proposition 4.5. The proof of this result relies on finite-precision estimates for the errors in a number of basic procedures.

These estimates are collected in the next section. We are aware they do not make the most exciting part of the paper but it is a part we cannot do without.

3 Error Bounds

In this section we show bounds for the basic computations occurring in ALHVar. We will use these bounds in subsequent sections to show our main result.

3.1 Basic facts

We recall the basics of a floating-point arithmetic which idealizes the usual IEEE standard arithmetic. In contrast to the standard model (as in [10]) we adapt our exposition to complex arithmetic. This system is defined by a set $\mathbb{F} \subset \mathbb{Q}[i]$ containing 0 (the *floating-point complex numbers*), a transformation $r_u : \mathbb{C} \rightarrow \mathbb{F}$ (the *rounding map*), and a constant $u \in \mathbb{R}$ (the *round-off unit*) satisfying $0 < u < 1$. The properties we require for such a system are the following:

- (i) For any $x \in \mathbb{F}$, $r_u(x) = x$. In particular, $r_u(0) = 0$.
- (ii) For any $x \in \mathbb{C}$, $r_u(x) = x(1 + \delta)$ with $|\delta| \leq u$.
- (iii) For any $y \in \mathbb{F}$, the set $r_u^{-1}(y)$ is measurable in \mathbb{C} .

Property (iii) ensures that the measure $\nu_{\mathbb{F}}$ described in the Introduction is well defined. Because of the enumerability of $\mathbb{Q}[i]$, this measure can be seen as a discretization of the Gaussian in $\mathcal{H}_{(\mathbf{d})}$.

We also define on \mathbb{F} arithmetic operations following the classical scheme

$$x \tilde{\circ} y = r_u(x \circ y)$$

for any $x, y \in \mathbb{F}$ and $\circ \in \{+, -, \times, /\}$, so that

$$\tilde{\circ} : \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{F}.$$

The following is an immediate consequence of property (ii) above.

Proposition 3.1. *For any $x, y \in \mathbb{F}$ we have*

$$x \tilde{\circ} y = (x \circ y)(1 + \delta), \quad |\delta| \leq u. \quad \square$$

When combining many operations in floating-point arithmetic, quantities such as $\prod_{i=1}^n (1 + \delta_i)^{\rho_i}$ naturally appear. Our round-off analysis uses the notations and ideas in Chapter 3 of [10], from where we quote the following results:

Proposition 3.2. *If $|\delta_i| \leq u$, $\rho_i \in \{-1, 1\}$, and $nu < 1$, then*

$$\prod_{i=1}^n (1 + \delta_i)^{\rho_i} = 1 + \theta_n,$$

where

$$|\theta_n| \leq \gamma_n = \frac{nu}{1 - nu}. \quad \square$$

Proposition 3.3. *For any positive integer k such that $ku < 1$, let θ_k, θ_j be any quantities satisfying*

$$|\theta_k| \leq \gamma_k = \frac{ku}{1 - ku} \quad |\theta_j| \leq \gamma_j = \frac{ju}{1 - ju}.$$

The following relations hold.

1. $(1 + \theta_k)(1 + \theta_j) = 1 + \theta_{k+j}$ for some $|\theta_{k+j}| \leq \gamma_{k+j}$.

2.

$$\frac{1 + \theta_k}{1 + \theta_j} = \begin{cases} 1 + \theta_{k+j} & \text{if } j \leq k, \\ 1 + \theta_{k+2j} & \text{if } j > k. \end{cases}$$

for some $|\theta_{k+j}| \leq \gamma_{k+j}$ or some $|\theta_{k+2j}| \leq \gamma_{k+2j}$.

3. If $ku, ju \leq 1/2$, then $\gamma_k \gamma_j \leq \gamma_{\min\{k,j\}}$.

4. $i\gamma_k \leq \gamma_{ik}$.

5. $\gamma_k + u \leq \gamma_{k+1}$.

6. $\gamma_k + \gamma_j + \gamma_k \gamma_j \leq \gamma_{k+j}$. □

From now on, whenever we write an expression containing θ_k we mean that the same expression is true for some θ_k , with $|\theta_k| \leq \gamma_k$.

When computing an arithmetic expression q with a round-off algorithm, errors will accumulate and we will obtain another quantity which we will denote by $\mathbf{fl}(q)$. For a complex number, we write $\mathbf{Error}(q) = |q - \mathbf{fl}(q)|$; for vectors or matrices, $\mathbf{Error}(q)$ will denote the vector or matrix of coordinates $|q_\alpha - \mathbf{fl}(q_\alpha)|$, allowing us to choose various norms to estimate this error.

An example of round-off analysis which will be useful in what follows is given in the next proposition, the proof of which follows the lines of the proof of the real version of this result that can be found in Section 3.1 of [10].

Proposition 3.4. *There is a finite-precision algorithm which, with input $x, y \in \mathbb{C}^n$, computes the inner product of x and y . The computed value $\mathbf{fl}(\langle x, y \rangle)$ satisfies*

$$\mathbf{fl}(\langle x, y \rangle) = \langle x, y \rangle + \theta_{\lceil \log_2 n \rceil + 1} \sum_{i=1}^n |x_i \bar{y}_i|.$$

In particular, if $x = y$, the algorithm computes $\mathbf{fl}(\|x\|^2)$ satisfying

$$\mathbf{fl}(\|x\|^2) = \|x\|^2 (1 + \theta_{\lceil \log_2 n \rceil + 1}). \quad \square$$

We assume that, besides the four basic operations, we are allowed to compute basic trigonometric functions (such as \sin and \cos) and the square root with finite precision. That is, if \mathbf{op} denotes any of these two operators, we compute $\widetilde{\mathbf{op}}$ such that

$$\widetilde{\mathbf{op}}(x) = \mathbf{op}(x)(1 + \delta), |\delta| < u.$$

The following sensitivity results will help us to deal with errors in computing trigonometric functions.

Lemma 3.5. (i) Let $t, \theta \in \mathbb{R}$. Then

$$|\cos(t + \theta) - \cos t| \leq |\theta|;$$

$$|\sin(t + \theta) - \sin t| \leq |\theta|;$$

(ii) Given two reals a and e such that both a and $a + e$ are in the interval $[0, 0.8]$, one has

$$|\arcsin(a + e) - \arcsin(a)| \leq 2|e|, \text{ with } |v| \leq |e|.$$

PROOF.

(i) Observe that

$$|\cos(t + \theta) - \cos t| = 2 \left| \sin \left(t + \frac{\theta}{2} \right) \right| \left| \sin \frac{\theta}{2} \right| \leq 2 \left| \sin \frac{\theta}{2} \right| \leq |\theta|,$$

and analogously

$$|\sin(t + \theta) - \sin t| = 2 \left| \cos \left(t + \frac{\theta}{2} \right) \right| \left| \sin \frac{\theta}{2} \right| \leq |\theta|. \quad \square$$

(ii) Without loss of generality, let us suppose that $e > 0$.

From the intermediate value theorem, there exists a ξ in $[a, a + e]$ such that $\arcsin(a + e) = \arcsin(a) + e \arcsin'(\xi) = \arcsin(a) + e \frac{1}{\sqrt{1 - \xi^2}}$.

Since $\xi \in [a, a + e]$, $|\xi| \leq 0.8$ and thus $|\arcsin'(\xi)| \leq \frac{1}{\sqrt{1 - 0.8^2}} < 2$. \square

To avoid burdening ourselves with the consideration of multiplicative constants, we introduce a further notation. Computational errors in our context are functions of the integer parameters n, N and D as well as on the condition $\mu_{\text{norm}}(g, z)$ of the initial pair (g, z) . For any such function Φ , we will write

$$\llbracket \Phi \rrbracket := \theta_{\mathcal{O}(\Phi)}.$$

The next properties follow directly from the properties of the θ notation.

Proposition 3.6. Let Φ and Ψ be two real functions. The following relations hold:

1. $\llbracket \Phi \rrbracket + \llbracket \Psi \rrbracket = \llbracket \max(\Phi, \Psi) \rrbracket$.
2. $\llbracket \Phi \rrbracket \llbracket \Psi \rrbracket = \llbracket \max(\Phi, \Psi) \rrbracket$.
3. If $\Phi \geq 1$, $\Phi \llbracket \Psi \rrbracket = \llbracket \Phi \Psi \rrbracket$.

3.2 Bounding errors for elementary computations

We now begin showing bounds for the errors in the crucial steps of our algorithm. To avoid burdening the exposition we will do so only for the steps dominating the accumulation of errors and simply warn the reader of the minor steps we consider as exact.

We begin with the evaluation of the errors in computing α . Remark that we suppose $\alpha \leq \pi/2$ in the following lemma. This will be ensured by the computation of α at the beginning of ALHVar. If this quantity is more than $\pi/2$, we set $f = -f$, ensuring that $\alpha \leq \pi/2$. We neglect the errors in this operation, and thus suppose in the remainder that $\alpha \leq \pi/2$.

Lemma 3.7. *Given f and g in $S(\mathcal{H}_{(\mathbf{d})})$ such that $d_{\mathbb{S}}(f, g) \leq \pi/2$, one can compute $\alpha = d_{\mathbb{S}}(f, g)$ with finite precision such that*

$$\mathbf{fl}(\alpha) = \alpha(1 + \llbracket \log N \rrbracket).$$

PROOF. As remarked in (3), one can compute $\alpha = d_{\mathbb{S}}(f, g)$ as $\alpha = 2 \arcsin\left(\frac{\|f-g\|}{2}\right)$.

We can compute the norm $\|f - g\|$ similarly as the vector norm in Proposition 3.4. In the case of polynomials in $\mathcal{H}_{(\mathbf{d})}$, the sum is over N coefficients, and thus we prove similarly that $\mathbf{fl}(\|f - g\|^2) = \|f - g\|^2(1 + \theta_{\lceil \log N \rceil + 1})$. Since we supposed that we can compute square root with finite precision, we get

$$\mathbf{fl}(\|f - g\|) = \|f - g\|(1 + \theta_{\lceil \log N \rceil + 2}).$$

Remark that, since we supposed $d_{\mathbb{S}}(f, g) \leq \pi/2$ and $\|f\| = \|g\| = 1$, we have $\|f - g\|/2 \leq \sin(\pi/4) = 1/\sqrt{2} < 0.71$. We can suppose that u is small enough such that the term $\theta_{\lceil \log N \rceil + 2}$ is smaller than $0.8 - 0.71$, and thus such that $\mathbf{fl}(\|f - g\|/2)$ is also in $[0, 0.8]$. We can thus apply Lemma 3.5, and by supposing that we are able to compute the function \arcsin with finite precision, we conclude that we can compute $\alpha = 2 \arcsin\left(\frac{\|f-g\|}{2}\right)$ such that

$$\begin{aligned} \mathbf{fl}(\alpha) &= \left(2 \arcsin\left(\frac{\|f-g\|}{2}\right) + 2 \frac{\|f-g\|}{2} \theta_{\mathcal{O}(\log N)}\right) (1 + \theta_{\mathcal{O}(1)}) \\ &= 2 \arcsin\left(\frac{\|f-g\|}{2}\right) (1 + \llbracket \log N \rrbracket), \end{aligned}$$

the last line since $\left|\frac{\|f-g\|}{2}\right| \leq \left|\arcsin\left(\frac{\|f-g\|}{2}\right)\right|$. \square

Proposition 3.8. *Given $\tau \in \mathbb{R}_+$, f and g in $S(\mathcal{H}_{(\mathbf{d})})$ such that $d_{\mathbb{S}}(f, g) \leq \pi/2$, we can calculate $t(\tau)$ with finite precision such that*

$$\mathbf{fl}(t) = t(1 + \llbracket \log N \rrbracket).$$

PROOF. First of all, observe that

$$\begin{aligned} t(\tau) = \frac{1}{\sin \alpha \cot(\alpha \tau) - \cos(\alpha) + 1} &= \frac{\sin(\tau \alpha)}{\sin \alpha \cos(\tau \alpha) - \cos \alpha \sin(\tau \alpha) + \sin(\tau \alpha)} \\ &= \frac{\sin(\tau \alpha)}{\sin(\alpha - \tau \alpha) + \sin(\tau \alpha)} \\ &= \frac{\sin(\tau \alpha)}{2 \sin\left(\frac{(1-\tau)\alpha + \tau \alpha}{2}\right) \cos\left(\frac{(1-\tau)\alpha - \tau \alpha}{2}\right)} \\ &= \frac{\sin(\tau \alpha)}{2 \sin \frac{\alpha}{2} \cos\left(\left(\frac{1}{2} - \tau\right) \alpha\right)}. \end{aligned}$$

We compute $t(\tau)$ via the last equality. First, we compute α following Lemma 3.7. Then, we show easily using Lemma 3.5 that each term in the fraction can be computed with finite precision up to a multiplicative factor $(1 + \theta_{\mathcal{O}(\log N)})$. We conclude using Proposition 3.3. \square

The following lemma bounds by $\|q\|$ the value of a polynomial q at any point on the unit sphere.

Lemma 3.9. *Given $d \in \mathbb{N}$, $q \in \mathbb{C}[X_0, \dots, X_n]$ homogeneous of degree d and $x \in S(\mathbb{C}^{n+1})$, we have $|q(x)| \leq \|q\|$.*

PROOF. Since our norm $\|\cdot\|$ on $\mathbb{C}[X_0, \dots, X_n]$ is unitarily invariant, for each element $\phi \in \mathcal{U}(n+1)$, one has $\|f \circ \phi\| = \|f\|$.

Let $e_0 := (1, 0, \dots, 0) \in \mathbb{C}^{n+1}$. Taking ϕ such that $\phi(e_0) = x$, one has

$$|q(x)| = |(q \circ \phi \circ \phi^{-1})(x)| = |(q \circ \phi)(e_0)|.$$

But $|q \circ \phi(e_0)|$ is exactly the coefficient of X_0^d in $q \circ \phi$ with respect to the Bombieri-Weyl basis of $\mathbb{C}[X_0, \dots, X_n]$, and thus $|q \circ \phi(e_0)| \leq \|q \circ \phi\| = \|q\|$. \square

Proposition 3.10. *Given $q \in S(\mathcal{H}_{\mathbf{d}})$ and $x \in S(\mathbb{C}^{n+1})$, we can compute $q(x)$ with finite precision u such that*

$$\|\text{Error}(q(x))\| = \llbracket \log N + D \rrbracket.$$

PROOF. For $i \leq n$, write $q_i(x) = \sum c_J x^J$. To compute $q_i(x)$ we compute each monomial $c_J x^J$ first, and then evaluate the sum. We have

$$\text{fl}(c_J x^J) = c_J x^J (1 + \theta_{d_i+1}),$$

and thus $\text{Error}(c_J x^J) \leq |c_J| |x|^J \gamma_{d_i+1}$.

As

$$\text{fl}(q_i(x)) = \text{fl}\left(\sum c_J x^J\right),$$

using pairwise summation (see section 4.2 in [10]) we have

$$\begin{aligned} \text{Error}(q_i(x)) &= \left| \sum \text{fl}(c_J x^J) - \sum (c_J x^J) \right| \\ &\leq \sum \text{Error}(c_J x^J) + \sum |c_J x^J| \gamma_{\lceil \log_2 N \rceil} \\ &\leq \sum |c_J| |x|^J (\gamma_{D+1} + \gamma_{\lceil \log_2 N \rceil} + \gamma_{D+1} \gamma_{\lceil \log_2 N \rceil}) \\ &\leq \sum |c_J| |x|^J \gamma_{\lceil \log_2 N \rceil + D+1}. \quad (\text{by Proposition 3.3 6.}) \end{aligned}$$

Note that $\sum |c_J| |x|^J \leq \|q_i\|$, by applying Lemma 3.9 to the polynomial of coefficients $|c_J|$, which has the same norm as q_i , at the point $|x| \in S(\mathbb{C}^{n+1})$. Hence,

$$\text{Error}(q_i(x)) \leq \|q_i\| \gamma_{\lceil \log_2 N \rceil + D+1},$$

and therefore,

$$\|\text{Error}(q(x))\|^2 \leq \gamma_{\lceil \log_2 N \rceil + D+1}^2 \sum_i \|q_i\|^2 = \gamma_{\lceil \log_2 N \rceil + D+1}^2 \|q\|^2 = \gamma_{\lceil \log_2 N \rceil + D+1}^2.$$

We finally have

$$\|\text{Error}(q(x))\| = \llbracket \log N + D \rrbracket. \quad \square$$

3.3 Bounding the error in the computation of $\mu_{\text{norm}}^{-1}(q, x)$

The bounds in **Error** ($\mu_{\text{norm}}^{-1}(q, x)$) scale well with q . Hence, to simplify notation, in all what follows we assume $\|q\| = 1$.

The main result in this subsection is the following.

Proposition 3.11. *Given $q \in S(\mathcal{H}_{(\mathbf{d})})$ and $x \in S(\mathbb{C}^{n+1})$ we can compute $\mu_{\text{norm}}^{-1}(q, x)$ satisfying*

$$\mathbf{Error}(\mu_{\text{norm}}^{-1}(q, x)) = \llbracket n(\log N + D + n) \rrbracket.$$

Note that under the assumption $\|q\| = 1$ our condition number becomes

$$\mu_{\text{norm}}(q, x) := \left\| (Dq(x)|_{T_x})^{-1} \mathbf{diag}(\sqrt{d_1}, \dots, \sqrt{d_n}) \right\|.$$

Given $q \in S(\mathcal{H}_{(\mathbf{d})})$ and $x \in S(\mathbb{C}^{n+1})$, let $M_q \in \mathbb{C}^{n \times n}$ be a matrix representing the linear operator

$$\begin{bmatrix} \frac{1}{\sqrt{d_1}} & & & \\ & \frac{1}{\sqrt{d_2}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{d_n}} \end{bmatrix} Dq(x)|_{T_x} \quad (5)$$

in some orthonormal basis of T_x (note that M_q depends also on x ; that point x will always be clear from the context). We then have $\mu_{\text{norm}}^{-1}(q, x) = \|M_q^{-1}\|^{-1} = \sigma_{\min}(M_q)$ where σ_{\min} denotes smallest singular value. We will compute $\mu_{\text{norm}}^{-1}(q, x)$ by computing M_q and then $\sigma_{\min}(M_q)$.

The following proposition contains several technical ideas that will help us to deal with the matrices $Dq(x)|_{T_x}$ and M_q . We use ideas from the proof of [7] modifying them to the complex case.

Proposition 3.12. *Let $q \in \mathcal{H}_{(\mathbf{d})}$ and $x \in S(\mathbb{C}^{n+1})$. Then the following statements are true:*

- (i) *The restriction of the derivative of q to the tangent space T_x can be represented by the following matrix :*

$$Dq(x)|_{T_x} = Dq(x)H,$$

where $H \in \mathbb{C}^{(n+1) \times n}$ is the matrix made with the last n columns of the matrix H_x defined by

$$H_x = \alpha(I_{n+1} - 2yy^*), \quad y = \frac{x - \alpha e_0}{\|x - \alpha e_0\|}, \quad \alpha = \frac{x_0}{|x_0|}$$

if $|x_0| \neq 1$, and $H_x = \alpha I_{n+1}$ otherwise.

- (ii)

$$\left\| \mathbf{diag} \left(\frac{1}{\sqrt{d_1}}, \dots, \frac{1}{\sqrt{d_n}} \right) Dq(x)|_{T_x} \right\| \leq \|q\|.$$

- (iii)

$$\|Dq(x)\|_F \leq \sqrt{D}\|q\|, \quad \|Dq(x)|_{T_x}\|_F \leq \sqrt{D}\|q\|.$$

PROOF. (i) For any unitary matrix H_x such that $H_x e_0 = x$, the n last columns H of H_x form an orthonormal basis of T_x . Thus $Dq(x)H$ is the representation of $Dq(x)|_{T_x}$ in that basis.

The matrix H_x computed here is constructed in [4]; one checks easily that it is unitary and that it satisfies $H_x e_0 = x$.

(ii) Let $g = q \circ H_x$. Then, differentiating the equality $g_i(H_x^* x) = q_i(x)$ and multiplying both sides by H on the right, we have

$$Dg_i(e_0)H_x^* H = Dq_i(x)H = Dq_i(x)|_{T_x}, \quad (6)$$

where the last equality is by (i). Observe that $H_x^* H = [e_1, \dots, e_n]$, hence,

$$Dg_i(e_0)H_x^* H = Dg_i(e_0)|_{T_{e_0}} = \left[\frac{\partial g_i}{\partial X_1}(e_0), \dots, \frac{\partial g_i}{\partial X_n}(e_0) \right]. \quad (7)$$

If we denote $g_i(X) = \sum_{\alpha} \binom{d}{\alpha}^{1/2} g_{i\alpha} X^{\alpha}$, it is straightforward that

$$\frac{\partial g_i}{\partial X_j}(e_0) = \binom{d_i}{d_i - 1}^{1/2} g_{i(e_j + (d_i - 1)e_0)} = \sqrt{d_i} \cdot g_{i(e_j + (d_i - 1)e_0)}.$$

Therefore, from (7),

$$\left\| \frac{1}{\sqrt{d_i}} Dg_i(e_0)|_{T_{e_0}} \right\|^2 = \sum_j g_{i(e_j + (d_i - 1)e_0)}^2 \leq \|g_i\|^2, \quad (8)$$

and hence by (8) we have

$$\begin{aligned} \left\| \text{diag} \left(\frac{1}{\sqrt{d_1}}, \dots, \frac{1}{\sqrt{d_n}} \right) Dg(e_0)|_{T_{e_0}} \right\|_F^2 &\leq \sum_{i=1}^n \left\| \frac{1}{\sqrt{d_i}} Dg_i(e_0)|_{T_{e_0}} \right\|^2 \\ &\leq \sum \|g_i\|^2 = \|g\|^2. \end{aligned} \quad (9)$$

Since the Hermitian inner product associated with the Bombieri-Weyl basis is unitarily invariant, we have

$$\|g\|^2 = \langle g, g \rangle = \langle q \circ H_x, q \circ H_x \rangle = \langle q, q \rangle = \|q\|^2,$$

which by (6), (7) and (9), and since the spectral norm of a matrix is not greater than its Frobenius norm, yields

$$\left\| \text{diag} \left(\frac{1}{\sqrt{d_1}}, \dots, \frac{1}{\sqrt{d_n}} \right) Dq(x)|_{T_x} \right\| = \left\| \text{diag} \left(\frac{1}{\sqrt{d_1}}, \dots, \frac{1}{\sqrt{d_n}} \right) Dg(e_0)|_{T_{e_0}} \right\| \leq \|g\| = \|q\|.$$

The relations (iii) can be shown similarly. \square

The following two statements are similar to those proved in [7] in the real case and similar ideas are used in the proofs.

Proposition 3.13. *Given $q \in S(\mathcal{H}_{(\mathbf{d})})$ and $x \in S(\mathbb{C}^{n+1})$, we have $\|Dq(x)|_{T_x}\| \leq \sqrt{D}$, and we can compute $Dq(x)|_{T_x}$ with finite precision such that*

$$\|\text{Error}(Dq(x)|_{T_x})\|_F \leq \llbracket n\sqrt{D}(\log N + D + \log n) \rrbracket.$$

PROOF. The inequality $\|Dq(x)|_{T_x}\| \leq \sqrt{D}$ follows from $\|q\| = 1$ and Proposition 3.12(iii).

We compute $Dq(x)|_{T_x}$ as in Proposition 3.12(i). Hence each entry (i, j) of the matrix $Dq(x)|_{T_x}$ is calculated as the product of $Dq_i(x)$ and the j th column $H_j = (h_{kj})_{1 \leq k \leq n+1}$ of H . Proceeding as in the proof of Proposition 3.10 we can compute $\frac{\partial q_i}{\partial X_k}(x)$ with

$$\mathbf{Error} \left(\frac{\partial q_i}{\partial X_k}(x) \right) = \llbracket (\log N + d_i) \rrbracket \|Dq_i(x)\|_F.$$

One can compute α as $\frac{x_0}{\sqrt{x_0 x_0^*}}$ with two arithmetic operations and one square root. Observe that to compute $x - \alpha e_0$, we need to perform only two more arithmetic operations. Also,

$$(yy^*)_{ij} = \frac{1}{\|x - \alpha e_0\|^2} ((x - \alpha e_0)(x - \alpha e_0)^*)_{ij}$$

and we have

$$\begin{aligned} \mathbf{fl}((x - \alpha e_0)(x - \alpha e_0)^*)_{ij} &= \mathbf{fl} \left((x - \alpha e_0)_i \overline{(x - \alpha e_0)_j} \right) \\ &= ((x - \alpha e_0)(x - \alpha e_0)^*)_{ij} (1 + \theta_{11}). \end{aligned}$$

Further,

$$\begin{aligned} \mathbf{fl}(\|x - \alpha e_0\|^2) &= \mathbf{fl} \left(\sum_{i=1}^n x_i \overline{x_i} + (x_0 - \alpha)(\overline{x_0} - \alpha) \right) \\ &= \|x - e_0\|^2 (1 + \theta_{\lceil \log_2(n+1) \rceil + 11}). \end{aligned}$$

Here we used pairwise summation bounds again.

Thus, by Proposition 3.3(2),

$$\mathbf{fl}(2yy^*)_{ij} = (2yy^*)_{ij} (1 + \theta_{2\lceil \log_2(n+1) \rceil + 35}).$$

Finally, taking into account one more addition and the multiplication by α , we get

$$\mathbf{Error}(h_{ij}) = \theta_{2\lceil \log_2(n+1) \rceil + 39} = \llbracket \log n \rrbracket.$$

Applying Proposition 3.4, we conclude

$$\begin{aligned} \mathbf{Error}([Dq(x)|_{T_x}]_{ij}) &= |\mathbf{fl}(\langle Dq_i(x), H_j \rangle) - \langle Dq_i(x), H_j \rangle| \\ &= |\langle \mathbf{fl}(Dq_i(x)), \mathbf{fl}(H_j) \rangle| \\ &\quad + \theta_{\lceil \log_2 n \rceil + 1} \sum_k |Dq_i(x)_k \overline{H_{kj}}| - \langle Dq_i(x), H_j \rangle| \\ &\leq |\langle \mathbf{Error}(Dq_i(x)), H_j \rangle| + |\langle Dq_i(x), \mathbf{Error}(H_j) \rangle| \\ &\quad + |\langle \mathbf{Error}(Dq_i(x)), \mathbf{Error}(H_j) \rangle| + \gamma_{\lceil \log_2 n \rceil + 1} |Dq_i(x)| |H_j| \\ &= (\llbracket (\log N + D) \rrbracket + \llbracket \sqrt{n} \log n \rrbracket \\ &\quad + \llbracket (\log N + D) \rrbracket \llbracket \sqrt{n} \log n \rrbracket + \llbracket \log n \rrbracket) \|Dq_i(x)\|_F \\ &= \llbracket \sqrt{n}(\log n + \log N + D) \rrbracket \|Dq_i(x)\|_F. \end{aligned}$$

This implies

$$\|\mathbf{Error}(Dq(x)|_{T_x})\|_F = \llbracket n(\log n + D + \log N) \rrbracket \|Dq(x)\|_F = \llbracket n\sqrt{D}(\log n + D + \log N) \rrbracket \quad \square$$

Proposition 3.14. *Given $q \in S(\mathcal{H}_{(\mathbf{d})})$, $x \in S(\mathbb{C}^{n+1})$ and M_q defined by (5), we have $\|M_q\| \leq 1$. In addition, we can compute such a matrix M_q with finite precision u such that*

$$\|\text{Error}(M_q)\|_F = \llbracket n(\log N + D + \log n) \rrbracket.$$

PROOF. The inequality $\|M_q\| \leq 1$ follows directly from Proposition 3.12(ii), as $\|q\| = 1$. Floating-point errors can be evaluated exactly as in Proposition 3.13; however, one gets rid of the factors \sqrt{D} since the bound on $\|M_q\|$ is better than the bound on $\|Dq(x)_{T_x}\|$. As a counterpart, one has to take into account one more division by $\sqrt{d_i}$ of each entry of the matrix, which slightly changes the constants, but leaves the order in N, D and n unchanged. \square

PROOF OF PROPOSITION 3.11. We use ideas from the proof of an analogous proposition in [7]. Let $x \in S(\mathbb{C}^{n+1})$, $q \in S(\mathcal{H}_{(\mathbf{d})})$ and M_q be as in Proposition 3.14. Then $\mu_{\text{norm}}^{-1}(q, x) = \sigma_{\min}(M_q) = \|M_q^{-1}\|^{-1}$ and we can compute the first expression by computing the last.

Let $E' = M_q - \text{fl}(M_q)$. By Proposition 3.14,

$$\|E'\| \leq \|E'\|_F = \llbracket n(\log N + D + \log n) \rrbracket.$$

Let $\mathcal{M}_q = \text{fl}(M_q)$. We compute $\sigma_{\min}(\mathcal{M}_q) = \|M_q^{-1}\|^{-1}$ using a backward stable algorithm (e.g., QR factorization). Then the computed $\text{fl}(\sigma_{\min}(\mathcal{M}_q))$ is the exact $\sigma_{\min}(\mathcal{M}_q + E'')$ for a matrix E'' with

$$\|E''\| \leq cn^2u\|\mathcal{M}_q\|$$

for some universal constant c (see, e.g., [9, 10]). Thus,

$$\text{fl}(\sigma_{\min}(M_q)) = \text{fl}(\sigma_{\min}(\mathcal{M}_q)) = \sigma_{\min}(\mathcal{M}_q + E'') = \sigma_{\min}(M_q + E' + E'').$$

Write $E = E' + E''$. Then, using $\|M_q\| \leq 1$ (by Proposition 3.14),

$$\begin{aligned} \|E\| &\leq \|E'\| + \|E''\| \leq \|E'\| + cn^2u\|\mathcal{M}_q\| \leq \|E'\| + cn^2u(\|M_q\| + \|E'\|) \\ &= \llbracket n(\log N + D + \log n) \rrbracket + \llbracket n^2 \rrbracket(1 + \llbracket n(\log N + D + \log n) \rrbracket) \\ &= \llbracket n(\log N + D + \log n) \rrbracket + \llbracket n(\log N + D + n) \rrbracket \\ &= \llbracket n(\log N + D + n) \rrbracket, \end{aligned}$$

using Proposition 3.6 in the penultimate row.

Therefore, $\text{fl}(\sigma_{\min}(M_q)) = \sigma_{\min}(M_q + E)$ which implies by [9, Corollary 8.3.2]:

$$\text{Error}(\sigma_{\min}(M_q)) \leq \|E\| < \llbracket n(\log N + D + n) \rrbracket. \quad \square$$

3.4 Bounding the error on the Newton step

We next evaluate the error in the computation of a Newton step. Our result is the following.

Proposition 3.15. *There exists a universal constant $\mathbf{e} > 0$ such that given a system $q \in S(\mathcal{H}_{(\mathbf{d})})$ and a point $x \in S(\mathbb{C}^{n+1})$, if the precision u satisfies*

$$u \leq \frac{\mathbf{e}}{D^2\mu_{\text{norm}}^2(q, x)n(D + \log N + n^2)},$$

then the error $\text{Error}(N_q(x))$ satisfies

$$\frac{\|\text{Error}(N_q(x))\|}{\|N_q(x)\|} \leq \frac{C(1 - \varepsilon)}{2\pi(1 + \varepsilon)^2 D^{3/2} \mu_{\text{norm}}(q, x)},$$

where C and ε are the constants introduced in Proposition 2.3.

We compute $N_q(x) - x$ by solving the linear system

$$\begin{bmatrix} Dq(x) \\ x^* \end{bmatrix} y = \begin{pmatrix} q(x) \\ 0 \end{pmatrix}.$$

We denote $D_q = \begin{bmatrix} Dq(x) \\ x^* \end{bmatrix}$.

Recall the following result from [10, Chapter 7] (in fact, Theorem 7.2 therein applied to $f = b/\|b\|$ and $E = A/\|A\|$).

Lemma 3.16. *Given a linear system $Ax = b$, approximations A' of A and b' of b such that $\|A - A'\| \leq \epsilon$, $\|b - b'\| \leq \epsilon$ and $\epsilon\|A^{-1}\| < 1$, the solution x' of the perturbed system $A'x' = b'$ satisfies :*

$$\|x' - x\| \leq \frac{\epsilon\|A^{-1}\|}{1 - \epsilon\|A^{-1}\|}(1 + \|x\|). \quad \square$$

Furthermore, from [10, Chapter 18], the solution \hat{x} of a linear system $Ax = b$, where A is non-singular, computed with a QR factorization with finite precision satisfies

$$(A + \Delta A)\hat{x} = b + \Delta b, \quad (10)$$

where $|\Delta A| \leq n^2\gamma_{cn}G|A|$, $|\Delta b| \leq n^2\gamma_{cn}G|b|$, $\|G\|_F = 1$, $c \in \mathbb{R}$. Here, $|A|$ denotes the matrix with entries $|a_{ij}|$ and the same for $|\Delta A|$, $|b|$, and $|\Delta b|$.

Lemma 3.17. *Let $q \in S(\mathcal{H}_d)$ and $x \in S(\mathbb{C}^{n+1})$. We can compute $N_q(x)$ with finite precision such that*

$$\frac{\|\mathbf{Error}(N_q(x))\|}{\|N_q(x)\|} = \llbracket \mu_{\text{norm}}(q, x) \cdot n\sqrt{D}(D + \log N + n^2) \rrbracket.$$

PROOF. To simplify notations, in this proof, we write $q(x)$ instead of $\begin{pmatrix} q(x) \\ 0 \end{pmatrix}$. Let y denote our computed solution of $D_q y = q(x)$.

From (10), $\mathbf{fl}(y)$ is the solution of a system $(\Delta D_q + \mathbf{fl}(D_q))y = \Delta q(x) + \mathbf{fl}(q(x))$ with $\|\Delta D_q\| = \llbracket n^3\sqrt{D} \rrbracket$ and $\|\Delta q(x)\| = \llbracket n^3 \rrbracket$.

From Proposition 3.10, given $q \in S(\mathcal{H}_d)$ and $x \in S(\mathbb{C}^{n+1})$, we can compute $q(x)$ with finite precision u such that $\|\mathbf{Error}(q(x))\| = \llbracket D + \log N \rrbracket$. Obviously, $\|\mathbf{Error}(D_q)\|$ is not greater than the bound we computed for $\|\mathbf{Error}(Dq(x)H)\|$ in Proposition 3.13. Hence, $\mathbf{Error}(D_q) = \llbracket n\sqrt{D}(\log N + D + \log n) \rrbracket$.

Furthermore, from (10), $\|\Delta D_q\| = \llbracket n^3\sqrt{D} \rrbracket$ and $\|\Delta q(x)\| = \llbracket n^3 \rrbracket$.

Finally, both terms $\|\Delta D_q + \mathbf{Error}(D_q)\|$ and $\|\Delta q(x) + \mathbf{Error}(q(x))\|$ can be bounded by an expression $\epsilon = \llbracket n\sqrt{D}(\log N + D + n^2) \rrbracket$.

Thus, from Lemma 3.16, the error on y is bounded as

$$\begin{aligned} \mathbf{Error}(y) &\leq \frac{\epsilon\|D_q^{-1}\|}{1 - \epsilon\|D_q^{-1}\|}(1 + \|y\|) \\ &= (1 + \|y\|)\llbracket \mu_{\text{norm}}(q, x)n\sqrt{D}(D + \log N + n^2) \rrbracket, \end{aligned}$$

the last line since $\|D_q^{-1}\| = \mathcal{O}(\mu_{\text{norm}}(q, x))$ and $\frac{1}{1 - \gamma_k} \leq \gamma_{4k}$.

Since $N_q(x)$ belongs to the tangent space to the unit sphere T_x , $\|N_q(x)\| \geq 1$, and

$$1 + \|y\| = 1 + \|N_q(x) - x\| \leq 1 + \|N_q(x)\| + \|x\| \leq 3\|N_q(x)\|.$$

Hence,

$$\|\mathbf{Error}(y)\| = 3\|N_q(x)\| \ll \mu_{\text{norm}}(q, x) n \sqrt{D}(D + \log N + n^2).$$

Then, the computation of $N_q(x)$ from $y = N_q(x) - x$ is a simple addition and does not change the order of the errors. \square

The proof of Proposition 3.15 is now immediate.

3.5 Bounding the error for $\Delta\tau$

We evaluate here the errors in the computation of the quantity $\Delta\tau$, that is, the size of the current step in the homotopy.

Proposition 3.18. *For $x \in S(\mathbb{C}^{n+1})$, and $f, g, q \in S(\mathcal{H}_{(\mathbf{d})})$ such that $d_{\mathbb{S}}(f, g) \leq \pi/2$ define the quantity*

$$\Delta\tau := \frac{\lambda}{d_{\mathbb{S}}(f, g) D^{3/2} \mu_{\text{norm}}^2(q, x)}.$$

There exists a universal constant $\mathbf{f} > 0$ such that

$$u \leq \frac{\mathbf{f}}{n(\log N + D + n) \mu_{\text{norm}}^2(q, x)} \quad (11)$$

implies

$$\mathbf{Error}(\Delta\tau) \leq \frac{1}{4} \Delta\tau.$$

To prove this proposition we rely on the following lemma.

Lemma 3.19. *Given $x \in S(\mathbb{C}^{n+1})$ and $q \in S(\mathcal{H}_{(\mathbf{d})})$, one can compute $\sigma_{\min}^2(M_q)$ with finite precision u such that*

$$\mathbf{Error}(\sigma_{\min}^2(M_q)) = \ll n(\log N + D + n).$$

PROOF. By Proposition 3.11, $\mathbf{Error}(\sigma_{\min}(M_q)) = \ll n(\log N + D + n)$. Hence, we have

$$\begin{aligned} |\mathbf{fl}(\sigma_{\min}^2(M_q)) - \sigma_{\min}^2(M_q)| &\leq 2|\sigma_{\min}(M_q)| \ll n(\log N + D + n) + \ll n(\log N + D + n)^2 \\ &\leq \ll n(\log N + D + n) + \ll n(\log N + D + n), \end{aligned}$$

since, by Proposition 3.14, $|\sigma_{\min}(M_q)| \leq \|M_q\| \leq 1$. Thus,

$$\mathbf{Error}(\sigma_{\min}^2(M_q)) = \ll n(\log N + D + n). \quad \square$$

PROOF OF PROPOSITION 3.18. One has

$$\begin{aligned} \mathbf{fl}(\Delta\tau) &= \mathbf{fl}\left(\frac{\lambda}{\alpha D^{3/2} \mu_{\text{norm}}^2(q, x)}\right) \\ &= \frac{\lambda}{D^{3/2}} \mathbf{fl}\left(\frac{\sigma_{\min}^2(M_q)}{\alpha}\right) (1 + \theta_{\mathcal{O}(1)}) \\ &= \frac{\lambda \mathbf{fl}(\sigma_{\min}^2(M_q))}{\alpha D^{3/2}} (1 + \theta_{\mathcal{O}(\log N)}), \end{aligned}$$

the last equality being from Lemma 3.7, and thus by Lemma 3.19

$$\begin{aligned}\text{Error}(\Delta\tau) &= \frac{\lambda}{\alpha D^{3/2}}(\llbracket n(\log N + D + n) \rrbracket + \llbracket \log N \rrbracket) \\ &= \frac{\lambda}{\alpha D^{3/2}}\llbracket n(\log N + D + n) \rrbracket.\end{aligned}$$

If u satisfies (11) with a value of \mathbf{f} small enough, the term $\llbracket n(\log N + D + n) \rrbracket$ may be bounded by

$$\llbracket n(\log N + D + n) \rrbracket \leq \frac{1}{4\mu_{\text{norm}}^2(q, x)},$$

and consequently

$$\text{Error}(\Delta\tau) \leq \frac{\lambda}{4\alpha D^{3/2}\mu_{\text{norm}}^2(q, x)} = \frac{1}{4}\Delta\tau. \quad \square$$

3.6 Bounding the distance between \tilde{q}_τ and q_τ

We evaluate here the error in the computation of q_τ , given f, g , and τ .

Proposition 3.20. *There exists a universal constant \mathbf{g} such that the following holds. Let $f, g \in S(\mathcal{H}_{(\mathbf{d})})$ with $d_{\mathbb{S}}(f, g) \leq \frac{\pi}{2}(1 + 1/6)$ be given with roundoff error u . Let $\tau \in [0, 1]$. Then for all $A \in (0, 1)$,*

$$u \leq \frac{\mathbf{g} \cdot A}{\log N}$$

implies

$$\|\mathbf{fl}(q_\tau) - q_\tau\| \leq A.$$

We first bound the distance between the points $tf + (1-t)g$ and $t\mathbf{fl}(f) + (1-t)\mathbf{fl}(g)$, without taking into account the error in the computation of t .

Proposition 3.21. *Assume that $f, g, \tilde{f}, \tilde{g} \in S(\mathcal{H}_{(\mathbf{d})})$ are such that $d_{\mathbb{S}}(f, g) \leq \frac{\pi}{2}(1 + 1/60)$ and $\|f - \tilde{f}\| \leq 1/60$, $\|g - \tilde{g}\| \leq 1/60$. For $t \in [0, 1]$ define $q = tf + (1-t)g$ and $\tilde{q} = t\tilde{f} + (1-t)\tilde{g}$. Then*

$$d_{\mathbb{S}}(q, \tilde{q}) \leq 2 \max \left\{ \|f - \tilde{f}\|, \|g - \tilde{g}\| \right\}.$$

To prove Proposition 3.21 we rely on the following lemmas.

Lemma 3.22. *Let $f, g \in \mathcal{H}_{(\mathbf{d})}$ with $\|f\|, \|g\| \geq \alpha > 0$, $\|f - g\| \leq \beta$ with $\alpha \geq \beta/2$. Then*

$$\frac{|\langle f, g \rangle|}{\|f\|\|g\|} \geq 1 - \frac{\beta^2}{2\alpha^2}. \quad (12)$$

PROOF. Pick any $f, g \in \mathcal{H}_{(\mathbf{d})}$ with $\|f\|, \|g\| \geq \alpha$, $\|f - g\| \leq \beta$, denote $r = (f + g)/2$, and let $s \in \mathcal{H}_{(\mathbf{d})}$ be such that $\|s\| = \|f - g\|/2$, $s \perp r$. Then from the orthogonality of r and s we have

$$\|r + s\|^2 = \|r - s\|^2 = \|r\|^2 + \|s\|^2 = \frac{\|f\|^2 + \|g\|^2}{2} \geq \|f\|\|g\| \geq \alpha^2;$$

also,

$$\|(r + s) - (r - s)\| = 2\|s\| = \|f - g\| \leq \beta.$$

Therefore,

$$\frac{|\langle r+s, r-s \rangle|}{\|r+s\|\|r-s\|} \leq \frac{||r\|^2 - \|s\|^2|}{\|f\|\|g\|} = \frac{||f+g\|^2 - \|f-g\|^2|}{4\|f\|\|g\|} \leq \frac{|\langle f, g \rangle|}{\|f\|\|g\|}.$$

Since $\|r+s\| = \|r-s\|$, we have

$$\min_{\substack{\|f\|, \|g\| \geq \alpha \\ \|f-g\| \leq \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|} = \min_{\substack{\|f\| = \|g\| \geq \alpha \\ \|f-g\| \leq \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|}. \quad (13)$$

Now assume that $f, g \in \mathcal{H}_{(\mathbf{d})}$ with $\|f\| = \|g\| \geq \alpha$, $\|f-g\| \leq \beta$. Let

$$\begin{aligned} f' &= \frac{\beta}{2} \cdot \frac{f-g}{\|f-g\|} + \frac{\sqrt{4\alpha^2 - \beta^2}}{2} \cdot \frac{f+g}{\|f+g\|}, \\ g' &= \frac{\beta}{2} \cdot \frac{g-f}{\|f-g\|} + \frac{\sqrt{4\alpha^2 - \beta^2}}{2} \cdot \frac{f+g}{\|f+g\|}. \end{aligned}$$

It is not difficult to check that $\|f'\| = \|g'\| = \alpha$ and $\|f' - g'\| = \beta$. Moreover,

$$\frac{|\langle f', g' \rangle|}{\|f'\|\|g'\|} = 1 - \frac{\beta^2}{2\alpha^2} \leq \frac{\|f\|^2 + \|g\|^2}{2\|f\|\|g\|} - \frac{\|f-g\|^2}{2\|f\|\|g\|} \leq \frac{|\langle f, g \rangle|}{\|f\|\|g\|}.$$

Therefore,

$$\min_{\substack{\|f\| = \|g\| \geq \alpha \\ \|f-g\| \leq \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|} = \min_{\substack{\|f\| = \|g\| = \alpha \\ \|f-g\| = \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|}. \quad (14)$$

From (13) and (14) we have

$$\min_{\substack{\|f\|, \|g\| \geq \alpha \\ \|f-g\| \leq \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|} = \min_{\substack{\|f\| = \|g\| = \alpha \\ \|f-g\| = \beta}} \frac{|\langle f, g \rangle|}{\|f\|\|g\|} \geq 1 - \frac{\beta^2}{2\alpha^2},$$

which shows (12). \square

Lemma 3.23. *Let $f, g \in \mathcal{H}_{(\mathbf{d})}$ with $\|f-g\| \leq \min\{\|f\|, \|g\|\}$. Then*

$$d_{\mathbb{S}}(f, g) < \frac{2}{\sqrt{3}} \cdot \frac{\|f-g\|}{\min\{\|f\|, \|g\|\}}. \quad (15)$$

From Lemma 3.22 we have

$$\cos d_{\mathbb{S}}(f, g) = \frac{\langle f, g \rangle}{\|f\|\|g\|} \geq 1 - \frac{\beta^2}{2\alpha^2},$$

where $\beta = \|f-g\|$, $\alpha = \min\{\|f\|, \|g\|\}$. From the Taylor expansion for \cos we obtain

$$\cos d_{\mathbb{S}}(f, g) \leq 1 - \frac{d_{\mathbb{S}}^2(f, g)}{2} + \frac{d_{\mathbb{S}}^4(f, g)}{24},$$

therefore

$$\frac{d_{\mathbb{S}}^4(f, g)}{12} - d_{\mathbb{S}}^2(f, g) + \frac{\beta^2}{\alpha^2} \geq 0.$$

Solving the relevant quadratic equation for $d_{\mathbb{S}}^2(f, g)$, we have

$$d_{\mathbb{S}}^2(f, g) \in \left(-\infty, 6 \left(1 - \sqrt{1 - \frac{\beta^2}{3\alpha^2}} \right) \right] \cup \left[6 \left(1 + \sqrt{1 - \frac{\beta^2}{3\alpha^2}} \right), \infty \right). \quad (16)$$

By our assumption $\beta/\alpha \leq 1$, therefore,

$$6 \left(1 + \sqrt{1 - \frac{\beta^2}{3\alpha^2}} \right) > \pi^2,$$

and the interval on the right-hand side of (16) is irrelevant (as $d_{\mathbb{S}}(f, g) \leq \pi$). We have (using $\beta/\alpha \leq 1$ again)

$$d_{\mathbb{S}}^2(f, g) \leq 6 \left(1 - \sqrt{1 - \frac{\beta^2}{3\alpha^2}} \right) = \frac{2}{1 + \sqrt{1 - \frac{\beta^2}{3\alpha^2}}} \cdot \frac{\beta^2}{\alpha^2} < \frac{4\beta^2}{3\alpha^2},$$

which yields (15). \square

Lemma 3.24. *Let $f, g \in \mathcal{H}_{(\mathbf{d})}$, $\|g\| = \|f\| = 1$, and $d_{\mathbb{S}}(f, g) \leq \frac{\pi}{2}(1 + \delta)$. Then, given $t \in [0, 1]$, $q(t) = tf + (1 - t)g$ satisfies*

$$\|q(t)\| \geq \sqrt{1 - \frac{\pi^2(1 + \delta)^2}{16}}. \quad (17)$$

PROOF. Consider the function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ defined as follows:

$$\varphi(t) = \|q(t)\|^2 = \|g\|^2 + 2t\Re\langle g, f - g \rangle + t^2\|f - g\|^2.$$

Observe that $\min_{t \in \mathbb{R}} \varphi(t)$ is attained at

$$t^* = -\frac{2\Re\langle g, f - g \rangle}{2\|f - g\|^2} = \frac{\|g - f\|^2 + \|g\|^2 - \|f\|^2}{2\|f - g\|^2} = \frac{1}{2},$$

and $\varphi(t^*) = \frac{1}{4}\|f + g\|^2$. We then have

$$\|q(t)\|^2 \geq \frac{1}{4}\|f + g\|^2 = 1 - \frac{\|f - g\|^2}{4} \geq 1 - \frac{d_{\mathbb{S}}^2(f, g)}{4} \geq 1 - \frac{\pi^2(1 + \delta)^2}{16},$$

which gives us (17). \square

PROOF OF PROPOSITION 3.21. Observe that

$$\begin{aligned} \|q - \tilde{q}\| &= \|tf + (1 - t)g - t\tilde{f} - (1 - t)\tilde{g}\| \\ &\leq t\|f - \tilde{f}\| + (1 - t)\|g - \tilde{g}\| \\ &\leq \max\{\|f - \tilde{f}\|, \|g - \tilde{g}\|\} \leq \frac{1}{60}. \end{aligned}$$

From Lemma 3.24 applied with $\delta = \frac{1}{6}$ we have

$$\|q\| \geq \sqrt{1 - \frac{(1 + 1/60)^2}{4}} > \frac{3}{5},$$

and hence

$$\|\tilde{q}\| \geq \|q\| - \|q - \tilde{q}\| > \frac{3}{5} - 1/60 = \frac{7}{12}.$$

Now applying Lemma 3.23, we have

$$d_{\mathbb{S}}(q, \tilde{q}) \leq \frac{2}{\sqrt{3}} \cdot \frac{\|q - \tilde{q}\|}{\min\{\|q\|, \|\tilde{q}\|\}} \leq \frac{2}{\sqrt{3}} \cdot \frac{\max\{\|f - \tilde{f}\|, \|g - \tilde{g}\|\}}{\frac{7}{12}} \leq 2 \max\{\|f - \tilde{f}\|, \|g - \tilde{g}\|\}.$$

□

PROOF OF PROPOSITION 3.20. Let us denote $\tilde{f} = \mathbf{fl}(f)$, $\tilde{g} = \mathbf{fl}(g)$ and $\tilde{t} = \mathbf{fl}(t)$. Let \tilde{q}_τ denote $\mathbf{fl}(q_\tau)$ and \hat{q}_τ the system $t\tilde{f} + (1 - t)\tilde{g}$. By hypothesis, both $\|f - \tilde{f}\|$ and $\|g - \tilde{g}\|$ are not greater than u .

Thus, by Proposition 3.21, if $u \leq 1/60$,

$$\|\hat{q}_\tau - q_\tau\| \leq 2u.$$

From Proposition 3.8, $\tilde{t} = t(1 + \theta_{\mathcal{O}(\log N)})$. Thus, there exists a constant \mathbf{g} such that for all $A \in (0, 1)$, $u \leq \frac{\mathbf{g}A}{\log N}$ implies

$$\|\hat{q}_\tau - \tilde{q}_\tau\| \leq A/2.$$

Taking $\mathbf{g} \leq 1/60$, $u \leq \frac{\mathbf{g}A}{\log N}$ ensures

$$\|\tilde{q}_\tau - q_\tau\| \leq \|\tilde{q}_\tau - \hat{q}_\tau\| + \|\hat{q}_\tau - q_\tau\| \leq \frac{5}{6}A < A.$$

□

3.7 Estimates for u

Along our homotopy, the precision needed to guarantee correctness varies with the system q considered. In our variable-precision algorithm we will want to keep this precision at all times within the interval $[\frac{1}{2}\mathbf{B}(q, x), \mathbf{B}(q, x)]$ with

$$\mathbf{B}(q, x) := \frac{k_2}{nD^2(\log N + D + n^2)\mu_{\text{norm}}^2(q, x)}, \quad (18)$$

where k_2 is a universal positive constant (that will be specified in Definition 3.28).

Now, since q and x vary at each iteration, one has to update the precision as well. To do so one faces an obstacle. When computing u we actually obtain a quantity $\mathbf{fl}(u)$ which depends on the current precision and this current precision has been computed in the previous iteration. Proposition 3.25 below shows that this obstacle can be overcome.

Proposition 3.25. *If $u \leq (1 + \varepsilon)^6 \mathbf{B}(q, x)$ then*

$$\text{Error} \left(\frac{3}{4}\mathbf{B}(q, x) \right) \leq \frac{1}{4}\mathbf{B}(q, x).$$

In particular, when computing $u := \frac{3}{4}\mathbf{B}(q, x)$ the computed quantity satisfies $\mathbf{fl}(u) \in [\frac{1}{2}\mathbf{B}(q, x), \mathbf{B}(q, x)]$.

Towards the proof of the proposition above we define

$$B(q, x) := \frac{1}{nD^2(\log N + D + n^2)\mu_{\text{norm}}^2(q, x)}$$

so that $\mathbf{B}(q, x) = k_2 B(q, x)$. Our first lemma bounds the error in the computation of $B(q, x)$.

Lemma 3.26. *There exists a positive universal constant k_3 such that the following is true. Assume $u \leq \frac{1}{2k_3n(\log N + D + n)}$. Then for any $x \in S(\mathbb{C}^{n+1})$ and $q \in S(\mathcal{H}_{(\mathbf{d})})$, one can compute $B(q, x)$ such that*

$$\text{Error}(B(q, x)) \leq \frac{k_3 u}{D^2}.$$

PROOF. From Lemma 3.19, we can compute $\sigma_{\min}^2(M_q)$ with error $\ll n(\log N + D + n)$; we can compute $B(q, x)$ such that

$$\text{fl}(B(q, x)) = \frac{\text{fl}(\sigma_{\min}^2(M_q))}{nD^2(\log N + D + n^2)}(1 + \theta_6),$$

and thus

$$\text{Error}(B(q, x)) = \frac{\ll n(\log N + D + n^2)\rrbracket}{nD^2(\log N + D + n^2)}.$$

It follows that there exists a constant k_3 such that

$$\text{Error}(B(q, x)) = \frac{\theta_{k_3n(\log N + D + n^2)}}{nD^2(\log N + D + n^2)}.$$

Thus, when $u \leq \frac{1}{2k_3n(\log N + D + n^2)}$, the denominator in $\gamma_{k_3n(\log N + D + n)}$ is greater than $1/2$, and

$$\text{Error}(B(q, x)) \leq \frac{k_3 u}{D^2}. \quad \square$$

Corollary 3.27. *Let k_2 be a positive constant such that $k_2 \leq \frac{1}{4k_3(1+\varepsilon)^6}$. The condition $u \leq (1+\varepsilon)^6 k_2 B(q, x)$ ensures that*

$$\text{Error}\left(\frac{3}{4}k_2 B(q, x)\right) \leq \frac{1}{4}k_2 B(q, x).$$

PROOF. If u is less than or equal to $(1+\varepsilon)^6 B(q, x)$, since $\mu_{\text{norm}}(q, x)$ is always greater than 1, if we choose k_2 not greater than $\frac{1}{2k_3(1+\varepsilon)^6}$, u will be less than or equal to $\frac{1}{2k_3n(\log N + D + n^2)}$. Thus, one has

$$\text{Error}(B(q, x)) \leq \frac{k_3 u}{D^2} \leq \frac{k_2 k_3 (1+\varepsilon)^6}{D^2} B(q, x).$$

Taking $k_2 \leq \frac{1}{4k_3(1+\varepsilon)^6}$ one has $\text{Error}(\frac{3}{4}k_2 B(q, x)) \leq \text{Error}(k_2 B(q, x)) \leq \frac{1}{4}k_2 B(q, x)$. \square

We now have all the conditions that the constant k_2 must fulfill.

Definition 3.28. Let k_2 be a positive constant chosen small enough such that

- (i) $k_2 B(q, x)$ is smaller than the bound on u in Proposition 3.18,
- (ii) $k_2 B(q, x)(1+\varepsilon)^4$ is smaller than the bound on u in Proposition 3.15,
- (iii) $k_2 \leq \frac{\mathbf{g}C(1-\varepsilon)}{12(1+\varepsilon)^5}$, where \mathbf{g} is defined in Proposition 3.20.
- (iv) k_2 verifies the condition in Corollary 3.27.

The first three conditions ensure that the precision will be good enough for the computation of the values of $\Delta\tau$, of the Newton operator and of q_τ . The fourth condition is needed for the computation of $B(q, x)$ itself.

PROOF OF PROPOSITION 3.25. From (18), the bound $\mathbf{B}(q, x)$ equals $k_2 B(q, x)$ and the result now follows from Corollary 3.27. \square

4 Analysis of the Homotopy

We next describe with more detail our procedure **ALHVar** —**A**daptive **L**inear **H**omotopy with **F**inite precision— to follow the path $\{(q_\tau, \zeta_\tau) \mid \tau \in [0, 1]\}$.

All the certifications on an execution of **ALHVar** will be for inputs satisfying certain conditions. We thus define the notion of admissible input for **ALHVar**.

Definition 4.1. An *admissible input* for algorithm **ALHVar** consists of

- A function `read_inputf()`, that returns an approximation of a system $f \in S(\mathcal{H}_{(\mathbf{d})})$ with the current round-off unit. That is, the instruction `read_inputf()` returns a system f' such that the coefficients a'_α of the polynomials f'_i satisfy

$$|a'_\alpha - a_\alpha| \leq u|a_\alpha|,$$

where a_α is the coefficient of the monomial of the same degree α of f_i . In particular, this implies that

$$\|f - f'\| \leq u\|f\|.$$

Note that `read_inputf()` is not required to be computable.

- An auxiliary system $g \in S(\mathcal{H}_{(\mathbf{d})})$, supposed to be given exactly.
- An approximate zero $x \in S(\mathbb{C}^{n+1})$ of g satisfying

$$d_{\mathbb{P}}(\zeta, x) \leq \frac{C}{D^{3/2}\mu_{\text{norm}}(g, \zeta)}$$

for its associated zero ζ .

- An initial round-off unit $u \in \mathbb{R}_+$ such that

$$u \leq \mathbf{B}(g, x).$$

For clarity, we denote such a tuple (f, g, x, u) and we refer to it as an input to **ALHVar** even though f is not given directly and the precision u is not passed as a parameter (it is a global variable in **MDVar**).

Define $\lambda := \frac{2C(1-\varepsilon)}{5(1+\varepsilon)^4} \approx 5.37 \cdot 10^{-3}$.

Algorithm ALHVar
input (f, g, x)
 $\tilde{f} := \text{read_input}_f(\)$
if $d_{\mathbb{S}}(\tilde{f}, g) \geq \frac{\pi}{2}$ then $g := -g$
 $\tau := 0, \tilde{q}_\tau := g$
repeat
 $\Delta\tau := \frac{\lambda}{d_{\mathbb{S}}(\tilde{f}, g) D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_\tau, x)}$
 $\tau := \min\{1, \tau + \Delta\tau\}$
 $\tilde{f} := \text{read_input}_f(\)$
 $\tilde{q}_\tau := t(\tau)\tilde{f} + (1 - t(\tau))g$
 $\tilde{q}_\tau := \frac{\tilde{q}_\tau}{\|\tilde{q}_\tau\|}$
 $x := N_{\tilde{q}_\tau}(x)$
 $x := \frac{x}{\|x\|}$
 $u := \frac{3}{4}\mathbf{B}(\tilde{q}_\tau, x)$
until $\tau = 1$
RETURN x

Remark 4.2. The algorithm ALHVar is a finite-precision adaptation of the algorithm ALH in [6]. It has a slightly smaller stepsize parameter λ . By the parameter f given to ALHVar, we mean, that the algorithm is given as input the procedure `read_inputf` that returns finite precision approximations of f .

We may use ALHVar to define a finite precision version MDVar of MD.

Algorithm MDVar
input $f \in \mathcal{H}_{(\mathbf{d})}$
 $u := \frac{k_2}{nD^2(\log N + D + n^2)2(n+1)^D}$
run ALHVar **on input** (f, \overline{U}, z_1)

To a pair $f \in S(\mathcal{H}_{(\mathbf{d})})$ and $(g, \zeta) \in V_{\mathbb{P}}$ we associate the number

$$\mu_*(f, g, \zeta) := \max_{\tau \in [0, 1]} \mu_{\text{norm}}(q_\tau, \zeta_\tau).$$

Theorem 4.3. *Let (f, g, x, u) be an admissible input of ALHVar. Then:*

- (i) *If the algorithm ALHVar stops on input (f, g, x) , it returns an approximate zero of f .*
- (ii) *Assume ALHVar stops on input (f, g, x) . Then, the number of iterations $K(f, g, x)$ performed by ALHVar satisfies*

$$K(f, g, x) \leq B(f, g, \zeta) + B(-f, g, \zeta)$$

where

$$B(f, g, \zeta) := 408 d_{\mathbb{S}}(f, g) D^{3/2} \int_0^1 \mu_{\text{norm}}^2(q_\tau, \zeta_\tau) d\tau.$$

Consequently the number of performed arithmetic operations $\text{cost}_{\text{ALHVar}}(f, g, x)$ is bounded by

$$\text{cost}_{\text{ALHVar}}(f, g, x) \leq \mathcal{O}(N)(B(f, g, \zeta) + B(-f, g, \zeta)).$$

If ALHVar does not stop then either $B(f, g, \zeta)$ or $B(-f, g, \zeta)$ is unbounded, and either the segment $E_{g, f}$ or $E_{g, -f}$ intersects Σ .

(iii) Furthermore, the finest precision $u_*(f, g, x)$ required during the execution is bounded from below by

$$u_*(f, g, x) = \Omega\left(\frac{1}{nD^2(\log N + D + n^2)\mu_*^2(f, g, \zeta)}\right).$$

4.1 Bounding errors in the homotopy

We begin with a simple consequence of Proposition 2.3.

Proposition 4.4. *Assume $D \geq 2$. Let $p_0, p_1 \in S(\mathcal{H}_{(\mathbf{d})})$, let ζ be a zero of p_0 , and A a positive constant not greater than C such that*

$$d_{\mathbb{S}}(p_0, p_1) \leq \frac{A}{(1 + \varepsilon)D^{3/2}\mu_{\text{norm}}^2(p_0, \zeta)}.$$

Then the path E_{p_0, p_1} can be lifted to a path in $V_{\mathbb{P}}$ starting in (p_0, ζ) . In addition, the zero χ of p_1 in this lifting satisfies

$$d_{\mathbb{P}}(\zeta, \chi) \leq \frac{A}{D^{3/2}\mu_{\text{norm}}(p_1, \chi)}.$$

Finally, for all $p_{\tau} \in E_{p_0, p_1}$, if ζ_{τ} denotes the zero of p_{τ} in this lifting, we have

$$\frac{1}{1 + \varepsilon}\mu_{\text{norm}}(p_0, \zeta) \leq \mu_{\text{norm}}(p_{\tau}, \zeta_{\tau}) \leq (1 + \varepsilon)\mu_{\text{norm}}(p_0, \zeta).$$

PROOF. For each $\tau \in [0, 1]$, let p_{τ} be the point of the segment $[p_0, p_1]$ such that $d_{\mathbb{S}}(p_0, p_{\tau}) = \tau d_{\mathbb{S}}(p_0, p_1)$.

Let τ_* be such that $\int_0^{\tau_*} \mu_{\text{norm}}(p_{\tau}, \zeta_{\tau}) \|\dot{p}_{\tau}\| d\tau = \frac{A}{D^{3/2}\mu_{\text{norm}}(p_0, \zeta)}$, or $\tau_* = 1$, or the path E_{p_0, p_1} cannot be lifted to V beyond τ_* , whichever is the smallest. Then, for all $\tau \in [0, \tau_*]$, using that $\|\dot{\zeta}_{\tau}\| \leq \mu_{\text{norm}}(p_{\tau}, \zeta_{\tau}) \|\dot{p}_{\tau}\|$ (cf. [5, §12.3-12.4]) we have

$$\begin{aligned} d_{\mathbb{P}}(\zeta, \zeta_{\tau}) &\leq \int_0^{\tau} \|\dot{\zeta}_s\| ds \leq \int_0^{\tau_*} \mu_{\text{norm}}(p_s, \zeta_s) \|\dot{p}_s\| ds \\ &\leq \frac{A}{D^{3/2}\mu_{\text{norm}}(p_0, \zeta)}. \end{aligned}$$

It is therefore enough to show that $\tau_* = 1$. Suppose to the contrary, that $\tau_* < 1$.

Since $\mu_{\text{norm}}(p_{\tau}, \zeta_{\tau}) \geq 1$, for every τ ,

$$d_{\mathbb{S}}(p_0, p_{\tau}) \leq d_{\mathbb{S}}(p_0, p_1) \leq \frac{A}{D^{3/2}\mu_{\text{norm}}(p_0, \zeta)}.$$

Since $A \leq C$ the bounds on $d_{\mathbb{S}}(p_0, p_\tau)$ and $d_{\mathbb{P}}(\zeta, \zeta_\tau)$ allow us to apply Proposition 2.3 and to deduce, for all $\tau \in [0, \tau_*]$,

$$\frac{\mu_{\text{norm}}(p_0, \zeta)}{1 + \varepsilon} \leq \mu_{\text{norm}}(p_\tau, \zeta_\tau) \leq (1 + \varepsilon)\mu_{\text{norm}}(p_0, \zeta). \quad (19)$$

We have

$$\begin{aligned} \frac{A}{D^{3/2}\mu_{\text{norm}}(p_0, \zeta)} &= \int_0^{\tau_*} \mu_{\text{norm}}(p_\tau, \zeta_\tau) \|\dot{p}_\tau\| d\tau \quad (\text{by definition of } \tau_*) \\ &\leq (1 + \varepsilon)\mu_{\text{norm}}(p_0, \zeta) \int_0^{\tau_*} \|\dot{p}_\tau\| d\tau \quad (\text{by (19)}) \\ &= d_{\mathbb{S}}(p_0, p_{\tau_*})(1 + \varepsilon)\mu_{\text{norm}}(p_0, \zeta), \end{aligned}$$

and thus

$$d_{\mathbb{S}}(p_0, p_{\tau_*}) \geq \frac{A}{(1 + \varepsilon)D^{3/2}\mu_{\text{norm}}^2(p_0, \zeta)} \geq d_{\mathbb{S}}(p_0, p_1),$$

which leads to a contradiction with $\tau_* < 1$, and finishes the proof. \square

The next proposition puts together many of the results obtained thus far. The general idea for its proof closely follows [6, Theorem 3.1] (which in turn is a constructive version of the main result in [12]) making some room for errors.

Let (f, g, x, u) be an admissible input for algorithm **ALHVar**.

Let $0 = \overline{\tau}_0 < \overline{\tau}_1 < \overline{\tau}_2 < \dots$, $x = \overline{x}_0, \overline{x}_1, \overline{x}_2, \dots$, and $\overline{u}_0, \overline{u}_1, \overline{u}_2, \dots$, be the sequences of τ -values, points in $S(\mathbb{C}^{n+1})$ and precisions generated by the algorithm **ALHVar** on the admissible input (f, g, x, u) . Let \tilde{f}_i be the approximation of the input f on the i th iteration.

Let $E_{g,f}$ be the path with endpoints g and f . To simplify notation we write q_i instead of $q_{\overline{\tau}_i}$ and ζ_i instead of $\zeta_{\overline{\tau}_i}$. Similarly, we denote by \tilde{q}_i the computed approximation of q_i —that is, $\tilde{q}_i = \mathbf{f}1(t(\tau_i)\tilde{f}_i + (1 - t(\tau_i))g)$ —, by x_{i+1} the exact value of $N_{q_i}(\overline{x}_i)$, and by τ_{i+1} the exact value of $\overline{\tau}_i + \Delta\tau$.

Proposition 4.5. *Let (f, g, x, u) be an admissible input for **ALHVar**. Let k be the number of iterations of **ALHVar** on input (f, g, x, u) —that is, either $k = \infty$ or $\tau_k = 1$, $q_k = f$. With the notations above, for all $i \in \{0, \dots, k - 1\}$, the following inequalities are true:*

- (a) $d_{\mathbb{P}}(\overline{x}_i, \zeta_i) \leq \frac{C}{D^{3/2}\mu_{\text{norm}}(q_i, \zeta_i)}$
- (u) $\frac{\mathbf{B}(\tilde{q}_i, \overline{x}_i)}{2} \leq \overline{u}_i \leq \mathbf{B}(\tilde{q}_i, \overline{x}_i)$
- (x) $d_{\mathbb{S}}(q_i, \tilde{q}_i) \leq \frac{C(1 - \varepsilon)}{12(1 + \varepsilon)D^{3/2}\mu_{\text{norm}}(q_i, \zeta_i)}$
- (c) $d_{\mathbb{S}}(q_i, q_{i+1}) \leq \frac{(1 - \varepsilon)C}{2(1 + \varepsilon)D^{3/2}\mu_{\text{norm}}(q_i, \zeta_i)}$
- (d) q_{i+1} has a zero ζ_{i+1} such that $d_{\mathbb{P}}(\zeta_i, \zeta_{i+1}) \leq \frac{(1 - \varepsilon)C}{2(1 + \varepsilon)D^{3/2}\mu_{\text{norm}}(q_i, \zeta_i)}$
- (e) \tilde{q}_{i+1} has a zero $\tilde{\zeta}_{i+1}$ such that $d_{\mathbb{P}}(\overline{x}_i, \tilde{\zeta}_{i+1}) \leq \frac{C((1 + \varepsilon) + 7/12(1 - \varepsilon))}{D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}$

Inequalities (a), (u), and (x) hold for k as well in case $k < \infty$.

Proposition 4.5 puts together all the needed bounds to ensure the proper work of **ALHVar**. Statement **(a,i)** ensures that $\overline{x_i}$ is “close enough” to ζ_i . That is, $\overline{x_i}$ is not just an approximate zero of q_i , but also an approximate zero for polynomials in a certain neighborhood of q_i on $E_{g,f}$. Statements **(c,i)** and **(d,i)** show that (taking into account computational errors) our step along the homotopy is so small that the next polynomial q_{i+1} belongs to this neighborhood. We hence arrive at **(e,i)**, which essentially means that $\overline{x_i}$ is an approximate zero of q_{i+1} associated with ζ_{i+1} . Therefore, the Newton step (with computational errors accounted for) brings the next iterate $\overline{x_{i+1}}$ close enough to ζ_{i+1} to ensure that **(a,i+1)** holds again. Making sure that **(u)** holds on every iteration, we guarantee that computational errors are small enough to allow all the other steps of the proof (**(a)**, **(c)**, **(d)** and **(e)**) to be carried through.

PROOF OF PROPOSITION 4.5. We proceed by induction by showing, that **(a,i)**, **(u,i)** and **(x,i)** imply successively **(c,i)**, **(d,i)**, **(x,i+1)**, **(e,i)**, and finally **(a,i+1)** and **(u,i+1)**.

Inequalities **(a)** and **(u)**, for $i = 0$ hold by hypothesis, and **(x, 0)** is obvious since $\tilde{q}_0 = q_0 = g$.

This gives us the induction base. Assume now that **(a)**, **(u)** and **(x)** hold for some $i \leq k-1$.

We now show **(c,i)** and **(d,i)**.

Observe that together with **(a,i)** and **(x,i)**, Proposition 2.3 implies

$$\frac{\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i})}{(1+\varepsilon)} \leq \mu_{\text{norm}}(q_i, \zeta_i) \leq (1+\varepsilon)\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i}). \quad (20)$$

By **(u,i)** and Definition 3.28 our precision $\overline{u_i}$ satisfies (11) for the pair $(\tilde{q}_i, \overline{x_i})$. Therefore, by Proposition 3.18 and the definition of $\Delta\tau$ in **ALHVar** we have

$$\begin{aligned} \alpha(\overline{\tau_{i+1}} - \overline{\tau_i}) &\leq \alpha(\text{Error}(\Delta\tau) + \tau_{i+1} - \overline{\tau_i}) \\ &\leq \alpha \frac{5}{4} \Delta\tau \leq \frac{\lambda(1 + \frac{1}{4})}{D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})}. \end{aligned}$$

So, using (20) and since $\lambda := \frac{2C(1-\varepsilon)}{5(1+\varepsilon)^4}$, we obtain

$$d_{\mathbb{S}}(q_i, q_{i+1}) = \alpha(\overline{\tau_{i+1}} - \overline{\tau_i}) \leq \frac{C(1-\varepsilon)}{2(1+\varepsilon)^4 D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})} \leq \frac{C(1-\varepsilon)}{2(1+\varepsilon)^2 D^{3/2} \mu_{\text{norm}}^2(q_i, \zeta_i)}.$$

Since $\mu_{\text{norm}}(q_i, \zeta_i)$ is always greater than or equal to 1, **(c,i)** holds, and **(d,i)** is the direct consequence of Proposition 4.4 applied to (q_i, q_{i+1}) and ζ_i , with $A = \frac{C(1-\varepsilon)}{2(1+\varepsilon)}$.

This application of Proposition 4.4 furthermore ensures that, for all $\tau \in [\overline{\tau_i}, \overline{\tau_{i+1}}]$,

$$\frac{\mu_{\text{norm}}(q_i, \zeta_i)}{1+\varepsilon} \leq \mu_{\text{norm}}(q_\tau, \zeta_\tau) \leq (1+\varepsilon)\mu_{\text{norm}}(q_i, \zeta_i), \quad (21)$$

and, in particular,

$$\frac{\mu_{\text{norm}}(q_i, \zeta_i)}{1+\varepsilon} \leq \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1}) \leq (1+\varepsilon)\mu_{\text{norm}}(q_i, \zeta_i). \quad (22)$$

Since $u \leq \mathbf{B}(\tilde{q}_i, \bar{x}_i)$ and from Definition 3.28, we can apply Proposition 3.20 with $A = \frac{C(1-\varepsilon)}{12(1+\varepsilon)^5 D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_i, \bar{x}_i)}$ and we get

$$\begin{aligned} d_{\mathbb{S}}(q_{i+1}, \tilde{q}_{i+1}) &\leq \frac{C(1-\varepsilon)}{12(1+\varepsilon)^5 D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_i, \bar{x}_i)} \\ &\leq \frac{\frac{C(1-\varepsilon)}{12(1+\varepsilon)^3}}{D^{3/2} \mu_{\text{norm}}^2(q_i, \zeta_i)} \quad (\text{from (20)}) \end{aligned} \quad (23)$$

and, hence, using (22),

$$d_{\mathbb{S}}(q_{i+1}, \tilde{q}_{i+1}) \leq \frac{\frac{C(1-\varepsilon)}{12(1+\varepsilon)}}{D^{3/2} \mu_{\text{norm}}^2(q_{i+1}, \zeta_{i+1})}. \quad (24)$$

Since $\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1}) \geq 1$ this shows $(\mathbf{x}, i+1)$.

We can now use (\mathbf{x}, i) , (\mathbf{c}, i) , and (23) to bound $d_{\mathbb{S}}(\tilde{q}_i, \tilde{q}_{i+1})$ as follows,

$$\begin{aligned} d_{\mathbb{S}}(\tilde{q}_i, \tilde{q}_{i+1}) &\leq d_{\mathbb{S}}(\tilde{q}_i, q_i) + d_{\mathbb{S}}(q_i, q_{i+1}) + d_{\mathbb{S}}(q_{i+1}, \tilde{q}_{i+1}) \\ &\leq \frac{\frac{C(1-\varepsilon)}{12(1+\varepsilon)}}{D^{3/2} \mu_{\text{norm}}(q_i, \zeta_i)} + \frac{\frac{C(1-\varepsilon)}{2(1+\varepsilon)}}{D^{3/2} \mu_{\text{norm}}(q_i, \zeta_i)} + \frac{\frac{C(1-\varepsilon)}{12(1+\varepsilon)^3}}{D^{3/2} \mu_{\text{norm}}^2(q_i, \zeta_i)} \\ &< \frac{\frac{C(1-\varepsilon)}{(1+\varepsilon)}}{D^{3/2} \mu_{\text{norm}}(q_i, \zeta_i)} \\ &\leq \frac{C(1-\varepsilon)}{D^{3/2} \mu_{\text{norm}}(\tilde{q}_i, \bar{x}_i)} \end{aligned} \quad (25)$$

the third inequality using $\mu_{\text{norm}}(q_i, \zeta_i) \geq 1$ and the last from (20). We can similarly bound distances between zeros and their approximations. Indeed, using (24), Proposition 4.4 applied to $(q_{i+1}, \tilde{q}_{i+1})$ and ζ_{i+1} , with $A = \frac{C(1-\varepsilon)}{12}$, ensures the existence of a zero $\tilde{\zeta}_{i+1}$ of \tilde{q}_{i+1} such that

$$d_{\mathbb{P}}(\zeta_{i+1}, \tilde{\zeta}_{i+1}) \leq \frac{C(1-\varepsilon)}{12 D^{3/2} \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}. \quad (26)$$

Next we use the triangle inequality to obtain

$$\begin{aligned} d_{\mathbb{P}}(\bar{x}_i, \tilde{\zeta}_{i+1}) &\leq d_{\mathbb{P}}(\bar{x}_i, \zeta_i) + d_{\mathbb{P}}(\zeta_i, \zeta_{i+1}) + d_{\mathbb{P}}(\zeta_{i+1}, \tilde{\zeta}_{i+1}) \\ &\leq \frac{C \left(1 + \frac{1-\varepsilon}{2(1+\varepsilon)}\right)}{D^{3/2} \mu_{\text{norm}}(q_i, \zeta_i)} + \frac{C \frac{1-\varepsilon}{12}}{D^{3/2} \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})} \quad (\text{by } (\mathbf{a}, i), (\mathbf{d}, i) \text{ and (26)}) \\ &\leq \frac{C(1+\varepsilon + \frac{7}{12}(1-\varepsilon))}{D^{3/2} \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}, \quad (\text{by (22)}) \end{aligned}$$

which proves (\mathbf{e}, i) .

Note that $(\mathbf{x}, i+1)$ and (26), together with Proposition 2.3, imply that $\mu_{\text{norm}}(\tilde{q}_{i+1}, \tilde{\zeta}_{i+1}) \leq (1+\varepsilon) \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})$. Also, that we have $C(1+\varepsilon)(1+\varepsilon + \frac{7}{12}(1-\varepsilon)) \leq \nu_0 \approx 0.3542$ and hence $d_{\mathbb{P}}(\bar{x}_i, \tilde{\zeta}_{i+1}) \leq \frac{\nu_0}{D^{3/2} \mu_{\text{norm}}(\tilde{q}_{i+1}, \tilde{\zeta}_{i+1})}$. We can therefore use Theorem 2.2 to deduce that \bar{x}_i is an approximate zero of \tilde{q}_{i+1} associated with its zero $\tilde{\zeta}_{i+1}$. Therefore, $x_{i+1} = N_{\tilde{q}_{i+1}}(\bar{x}_i)$ satisfies

$$d_{\mathbb{P}}(x_{i+1}, \tilde{\zeta}_{i+1}) \leq \frac{1}{2} d_{\mathbb{P}}(\bar{x}_i, \tilde{\zeta}_{i+1}) \leq \frac{C(1+\varepsilon + \frac{7}{12}(1-\varepsilon))}{2 D^{3/2} \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}, \quad (27)$$

where the last inequality is due to (\mathbf{e}, i) , and thus

$$\begin{aligned}
d_{\mathbb{P}}(x_{i+1}, \zeta_{i+1}) &\leq d_{\mathbb{P}}(x_{i+1}, \tilde{\zeta}_{i+1}) + d_{\mathbb{P}}(\tilde{\zeta}_{i+1}, \zeta_{i+1}) \\
&\leq \frac{\frac{1}{2}C(1+\varepsilon + \frac{7}{12}(1-\varepsilon)) + \frac{1}{12}C(1-\varepsilon)}{D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})} \\
&= \frac{C(\frac{1}{2}(1+\varepsilon) + \frac{3}{8}(1-\varepsilon))}{D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}.
\end{aligned} \tag{28}$$

Now we are ready to prove the last two implications. We first show $(\mathbf{a}, i+1)$.

Inequality (25) allows us to use once more Proposition 2.3 to deduce

$$\frac{1}{1+\varepsilon}\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i}) \leq \mu_{\text{norm}}(\tilde{q}_{i+1}, \overline{x_i}) \leq (1+\varepsilon)\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i}). \tag{29}$$

Since \overline{u}_i is less than or equal to $\mathbf{B}(\tilde{q}_i, \overline{x_i})$ (by (\mathbf{u}, i)), from the choice of the constant k_2 in Definition 3.28(ii) one has

$$\begin{aligned}
\overline{u}_i &\leq \frac{\mathbf{e}}{(1+\varepsilon)^2 n D^2 (\log N + D + n^2) \mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})} \\
&\leq \frac{\mathbf{e}}{n D^2 (\log N + D + n^2) \mu_{\text{norm}}^2(\tilde{q}_{i+1}, \overline{x_i})} \quad (\text{by (29)}).
\end{aligned}$$

The condition on u (for the pair $(\tilde{q}_{i+1}, \overline{x_i})$) of Proposition 3.15 is thus verified, and applying this proposition we obtain

$$\|\overline{x_{i+1}} - x_{i+1}\| = \text{Error}(N_{\tilde{q}_{i+1}}(\overline{x_i})) \leq \frac{C(1-\varepsilon)}{4\pi(1+\varepsilon)^2 D^{3/2} \mu_{\text{norm}}(\tilde{q}_{i+1}, \overline{x_i})}. \tag{30}$$

The proof of (25) implicitly shows that $d_{\mathbb{S}}(q_i, \tilde{q}_{i+1}) \leq \frac{C(1-\varepsilon)}{D^{3/2} \mu_{\text{norm}}(q_i, \zeta_i)}$. Together with (\mathbf{a}, i) we are in the hypothesis of Proposition 2.3 and we can deduce

$$\frac{1}{1+\varepsilon}\mu_{\text{norm}}(q_i, \zeta_i) \leq \mu_{\text{norm}}(\tilde{q}_{i+1}, \overline{x_i}) \leq (1+\varepsilon)\mu_{\text{norm}}(q_i, \zeta_i).$$

This inequality, together with (22), yields

$$\frac{1}{(1+\varepsilon)^2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1}) \leq \mu_{\text{norm}}(\tilde{q}_{i+1}, \overline{x_i}) \leq (1+\varepsilon)^2\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1}) \tag{31}$$

and using these bounds (30) becomes

$$\|\overline{x_{i+1}} - x_{i+1}\| \leq \frac{C(1-\varepsilon)}{4\pi D^{3/2} \mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})}. \tag{32}$$

We now use this bound and the triangle inequality to bound $d_{\mathbb{P}}(\overline{x_{i+1}}, \zeta_{i+1})$ as follows

$$\begin{aligned}
d_{\mathbb{P}}(\overline{x_{i+1}}, \zeta_{i+1}) &\leq d_{\mathbb{P}}(\overline{x_{i+1}}, x_{i+1}) + d_{\mathbb{P}}(x_{i+1}, \zeta_{i+1}) \\
&\leq \frac{\pi}{2}\|\overline{x_{i+1}} - x_{i+1}\| + d_{\mathbb{P}}(x_{i+1}, \zeta_{i+1}) \\
&\leq \frac{C(1-\varepsilon)}{8D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})} + \frac{C(1+\varepsilon + 3/4(1-\varepsilon))}{2D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})} \quad (\text{by (32) and (28)}) \\
&= \frac{C(\frac{1}{2}(1+\varepsilon) + \frac{3}{8}(1-\varepsilon) + \frac{1}{8}(1-\varepsilon))}{D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})} = \frac{C}{D^{3/2}\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1})},
\end{aligned}$$

which proves **(a)** for $i + 1$.

It remains to show **(u, i + 1)**. To do so note that we may use **(a, i + 1)** and **(x, i + 1)** together with Proposition 2.3 to obtain (20) for $i + 1$ (just as we obtained it for i). Consequently,

$$\begin{aligned}\mu_{\text{norm}}(\tilde{q}_{i+1}, \overline{x_{i+1}}) &\leq (1 + \varepsilon)\mu_{\text{norm}}(q_{i+1}, \zeta_{i+1}) \\ &\leq (1 + \varepsilon)^2\mu_{\text{norm}}(q_i, \zeta_i) \quad (\text{by (22)}) \\ &\leq (1 + \varepsilon)^3\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i}) \quad (\text{by (20)}).\end{aligned}$$

Using this bound along with **(u, i)** we obtain

$$\begin{aligned}\overline{u}_i &\leq \mathbf{B}(\tilde{q}_i, \overline{x_i}) = \frac{k_2}{nD^2(\log N + D + n^2)\mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})} \\ &\leq \frac{k_2(1 + \varepsilon)^6}{nD^2(\log N + D + n^2)\mu_{\text{norm}}^2(\tilde{q}_{i+1}, \overline{x_{i+1}})} = (1 + \varepsilon)^6\mathbf{B}(\tilde{q}_{i+1}, \overline{x_{i+1}}).\end{aligned}$$

We can therefore apply Proposition 3.25 with the pair $(\tilde{q}_{i+1}, \overline{x_{i+1}})$ to deduce that $\text{Error}(\frac{3}{4}\mathbf{B}(\tilde{q}_{i+1}, \overline{x_{i+1}})) \leq \frac{1}{4}\mathbf{B}(\tilde{q}_{i+1}, \overline{x_{i+1}})$, and consequently

$$\left| \overline{u}_{i+1} - \frac{3}{4}\mathbf{B}(\tilde{q}_{i+1}, \overline{x_{i+1}}) \right| \leq \frac{1}{4}\mathbf{B}(\tilde{q}_{i+1}, \overline{x_{i+1}}),$$

which proves **(u, i + 1)**. □

4.2 Proof of Theorem 4.3

(i) Since (f, g, x, u) is an admissible input for ALHVar we can use Proposition 4.5 (and the notation therein). The estimate $d_{\mathbb{P}}(\overline{x_k}, \zeta_k) \leq \frac{C}{D^{3/2}\mu_{\text{norm}}(q_k, \zeta_k)}$ shown as **(a, k)** in that proposition implies by Theorem 2.2 that the returned point $\overline{x_k}$ is an approximate zero of $q_k = f$ with associated zero ζ_1 .

(ii) The first instruction in ALHVar swaps f by $-f$ if $d_{\mathbb{S}}(\tilde{f}, g) \geq \frac{\pi}{2}$. The reason to do so is that for nearly antipodal instances of f and g the difference $d_{\mathbb{S}}(f, \tilde{f})$ may be arbitrarily magnified in $d_{\mathbb{S}}(q_{\tau}, \tilde{q}_{\tau})$. This does not occur under the assumption of infinite precision and this is why such swap is not in the algorithms described in [4, 6].

Let h be either $-f$ or f (according to whether ALHVar did the swap or not), $K(h, g, x)$ be the number of iterations performed by ALHVar, and $\{(q_{\tau}, \zeta_{\tau})\}$ be the lifting of the path $E_{g, h}$.

Let $k \leq K(h, g, x)$ be a positive integer and consider any $i \in \{0, \dots, k - 1\}$. Using Proposition 4.4 for q_i, q_{i+1} together with (20) implies that, for all $\tau \in [\overline{\tau}_i, \overline{\tau}_{i+1}]$,

$$\frac{\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i})}{(1 + \varepsilon)^2} \leq \mu_{\text{norm}}(q_{\tau}, \zeta_{\tau}) \leq (1 + \varepsilon)^2\mu_{\text{norm}}(\tilde{q}_i, \overline{x_i}). \quad (33)$$

Therefore,

$$\begin{aligned}
\int_{\overline{\tau_i}}^{\overline{\tau_{i+1}}} \mu_{\text{norm}}^2(q_\tau, \zeta_\tau) d\tau &\geq \int_{\overline{\tau_i}}^{\overline{\tau_{i+1}}} \frac{\mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})}{(1+\varepsilon)^4} d\tau \\
&= \frac{\mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})}{(1+\varepsilon)^4} (\overline{\tau_{i+1}} - \overline{\tau_i}) \\
&\geq \frac{\mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})}{(1+\varepsilon)^4} \frac{3\lambda}{4\alpha D^{3/2} \mu_{\text{norm}}^2(\tilde{q}_i, \overline{x_i})} \quad (\text{by Proposition 3.18}) \\
&= \frac{3\lambda}{4(1+\varepsilon)^4 \alpha D^{3/2}}.
\end{aligned}$$

If $k = K(h, g, x) < \infty$ this implies

$$\int_0^1 \mu_{\text{norm}}^2(q_\tau, \zeta_\tau) d\tau \geq \left(\frac{3\lambda}{4(1+\varepsilon)^4} \right) k \frac{1}{\alpha D^{3/2}} \geq k \frac{1}{408 \alpha D^{3/2}},$$

which proves that

$$K(h, g, x) \leq 408 d_{\mathbb{S}}(h, g) D^{3/2} \int_0^1 \mu_{\text{norm}}^2(q_\tau, \zeta_\tau) d\tau = B(h, g, \zeta). \quad (34)$$

It follows that the number of iterations $K(f, g, x)$ satisfies either $K(f, g, x) \leq B(f, g, \zeta)$ or $K(f, g, x) \leq B(-f, g, \zeta)$. Certainly —and this introduces a factor of 2 but simplifies the exposition—

$$K(f, g, x) \leq B(f, g, \zeta) + B(-f, g, \zeta).$$

In case $K(h, g, x) = \infty$ (a non-halting computation) it implies that $\int_0^1 \mu_{\text{norm}}^2(q_\tau, \zeta_\tau) d\tau = \infty$.

The bound for $\text{cost}_{\text{ALHVar}}$ follows from the $\mathcal{O}(N)$ cost of each iteration of ALHVar mentioned in §2.2.

(iii) For $i = 1, \dots, k-1$, due to (u,i),

$$\overline{u}_i \geq \frac{\mathbf{B}(\tilde{q}_i, \overline{x_i})}{2} = \Omega \left(\frac{1}{nD^2(\log N + D + n^2) \max_{\tau \in [\overline{\tau_i}, \overline{\tau_{i+1}}]} \mu_{\text{norm}}^2(q_\tau, \zeta_\tau)} \right)$$

the last by (33). The statement now follows from the equalities

$$u_*(f, g, \zeta) = \min_{i < k} \overline{u}_i \quad \text{and} \quad \mu_*(f, g, \zeta) = \max_{i < k} \max_{\tau \in [\overline{\tau_i}, \overline{\tau_{i+1}}]} \mu_{\text{norm}}(q_\tau, \zeta_\tau). \quad \square$$

5 Proof of Theorem B

We follow here the proof of the corresponding result for MD in [6] and begin by recalling two facts from this article. The first estimates the mean square condition number on the path when an extremity is fixed.

Theorem 5.1 (Theorem 10.1 in [6]). *For $g \in S(\mathcal{H}_{(\mathbf{d})}) \setminus \Sigma$ we have*

$$\mathbb{E}_{f \in S(\mathcal{H}_{(\mathbf{d})})} \left(d_{\mathbb{S}}(f, g) \int_0^1 \mu_2^2(q_\tau) d\tau \right) \leq 818 D^{3/2} N(n+1) \mu_{\text{max}}^2(g) + 0.01. \quad \square$$

The second bounds the condition of \overline{U} .

Lemma 5.2 (Lemma 10.5 in [6]). *The maximum of the condition numbers $\mu_{\max}(\overline{U}) := \max_{\mathbf{z}: \overline{U}(\mathbf{z})=0} \{\mu_{\text{norm}}(\overline{U}, \mathbf{z})\}$ satisfies*

$$\mu_{\max}^2(\overline{U}) \leq 2n \max_{i \leq n} \frac{1}{d_i} (n+1)^{d_i-1} \leq 2(n+1)^D. \quad \square$$

The following proposition bounds the maximum $\mu_*(f, g, \zeta)$ of the condition number along a path from (g, ζ) to f in terms of the number of iterations of ALHVar to follow this path and of the condition number $\mu_{\text{norm}}(g, \zeta)$ of the initial pair.

Proposition 5.3. *Let $f, g \in S(\mathcal{H}_{(\mathbf{d})})$ and ζ a zero of g . The largest condition number $\mu_*(f, g, \zeta)$ along the path from (g, ζ) to f satisfies*

$$\mu_*(f, g, \zeta) \leq (1 + \varepsilon)^{K(f, g, \zeta)} \mu_{\text{norm}}(g, \zeta).$$

PROOF. Write $k := K(f, g, \zeta)$ and let $\mu_{*i} := \max_{\tau \in [\overline{\tau}_i, \overline{\tau}_i+1]} \mu_{\text{norm}}(q_\tau, \zeta_\tau)$. With this notation, we have $\mu_*(f, g, \zeta) = \max_{i=0, \dots, k-1} \mu_{*i}$. Furthermore, (21) states that, for all $i \leq k-1$,

$$\mu_{*i} \leq (1 + \varepsilon) \mu_{\text{norm}}(q_i, \zeta_i)$$

and an immediate recursion yields

$$\mu_*(f, g, \zeta) = \max_{i \in \{1, \dots, k-1\}} \mu_{*i} \leq (1 + \varepsilon)^k \mu_{\text{norm}}(g, \zeta). \quad \square$$

We remark that from the unitary invariance of our setting, for any unitary transformation $\nu \in \mathcal{U}(n+1)$ and any $g \in \mathcal{H}_{(\mathbf{d})}$ and $x \in \mathbb{P}^n$,

$$\mu_{\text{norm}}(g, x) = \mu_{\text{norm}}(g \circ \nu^{-1}, \nu x).$$

Furthermore, for any execution of ALHVar on an admissible input (f, g, x) , the number of iterations $K(f, g, x)$ during the execution satisfies $K(f, g, x) = K(f \circ \nu^{-1}, g \circ \nu^{-1}, \nu x)$ for any unitary transformation $\nu \in \mathcal{U}(n+1)$.

But one can remark also that any zero \mathbf{z}_i of \overline{U} is the image of $\mathbf{z}_1 = \frac{1}{\sqrt{2n}}(1, \dots, 1)$ by a unitary transformation ν_i that leaves \overline{U} invariant. Thus, $K(f, \overline{U}, \mathbf{z}_1) = K(f \circ \nu_j^{-1}, \overline{U}, \mathbf{z}_i)$ for all zeros \mathbf{z}_i of \overline{U} , and $\mu_{\max}(\overline{U}) = \mu_{\text{norm}}(\overline{U}, \mathbf{z}_1)$.

We also obtain immediately

$$K(f, \overline{U}, \mathbf{z}_1) = \frac{1}{D} \sum_{j=1}^D K(f \circ \nu_j^{-1}, \overline{U}, \mathbf{z}_j). \quad (35)$$

But for all measurable functions $\varphi: S(\mathcal{H}_{(\mathbf{d})}) \rightarrow \mathbb{R}$ and all $\nu \in \mathcal{U}(n+1)$ we have

$$\mathbb{E}_{f \in S(\mathcal{H}_{(\mathbf{d})})} \varphi(f) = \mathbb{E}_{f \in S(\mathcal{H}_{(\mathbf{d})})} \varphi(f \circ \nu),$$

due to the isotropy of the uniform measure on $S(\mathcal{H}_{(\mathbf{d})})$.

Therefore, (35) implies

$$\mathbb{E}_{f \in S(\mathcal{H}_{(d)})} K(f, \overline{U}, z_1) = \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} \frac{1}{D} \sum_{j=1}^D K(f, \overline{U}, z_j). \quad (36)$$

PROOF OF THEOREM B. From Lemma 5.2, $\mu_{\text{norm}}(\overline{U}, z_1) \leq \sqrt{2}(n+1)^{D/2}$, and thus the initial value for u in algorithm MDVar is less than or equal to $\mathbf{B}(\overline{U}, z_1)$. Therefore, the tuple $(f, \overline{U}, z_1, u)$ given to ALHVar during the execution of MD is an admissible input, and we can apply Theorem 4.3 to that execution of ALHVar. In particular, it follows that MDVar almost surely stops and when it does so, it returns an approximate zero of f .

We next bound the average cost of MDVar. Recall, we denoted by $K(f, \overline{U}, z_1)$ the number of iterations of ALHVar during the execution of MDVar with input f . Again by Theorem 4.3, for any root z_j of \overline{U} we have

$$K(f, \overline{U}, z_j) \leq B(f, \overline{U}, z_j) + B(-f, \overline{U}, z_j).$$

But we obviously have that $\mathbb{E}_{f \in S(\mathcal{H}_{(d)})} B(f, \overline{U}, z_j) = \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} B(-f, \overline{U}, z_j)$, and thus from (36),

$$\begin{aligned} \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} K(f, \overline{U}, z_1) &\leq 2 \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} \frac{1}{D} \sum_{j=1}^D B(f, \overline{U}, z_j) \\ &= 816D^{3/2} \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} d_{\mathbb{S}}(f, \overline{U}) \int_0^1 \frac{1}{D} \sum_{j=1}^D \mu_{\text{norm}}^2(q_\tau, \zeta_\tau^{(j)}) d\tau \\ &= 816D^{3/2} \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} d_{\mathbb{S}}(f, \overline{U}) \int_0^1 \mu_2^2(q_\tau) d\tau, \end{aligned}$$

the last line by the definition of the mean square condition number (2).

Applying successively Theorem 5.1 and Lemma 5.2, we get

$$\begin{aligned} \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} K(f, \overline{U}, z_1) &\leq 816D^{3/2}(818D^{3/2}N(n+1)\mu_{\max}^2(\overline{U}) + 0.01) \\ &\leq 816D^{3/2}(818D^{3/2}N(n+1) \cdot 2(n+1)^D + 0.01) \\ &= 667488D^3N(n+1)^{D+1} + 8.16D^{3/2} \\ &\leq 667489D^3N(n+1)^{D+1}. \end{aligned} \quad (37)$$

The bound for the average of $\text{cost}_{\text{MDVar}}$ follows from the $\mathcal{O}(N)$ cost of each iteration of ALHVar.

We finally bound the average of the precision needed. From Theorem 4.3 (iii), the finest precision $u_*(f, \overline{U}, z_1)$ along the execution of ALHVar (and therefore, along that of MDVar) satisfies, for some universal constant c ,

$$u_*(f, \overline{U}, z_1) \geq \frac{1}{cnD^2(\log N + D + n^2)\mu_*^2(f, \overline{U}, z_1)}.$$

Hence by Proposition 5.3,

$$\begin{aligned} |\log u_*(f, \overline{U}, z_1)| &\leq \log(cnD^2(\log N + D + n^2)\mu_*^2(f, \overline{U}, z_1)) \\ &= 2\log \mu_*(f, \overline{U}, z_1) + \log(cnD^2(\log N + D + n^2)) \\ &\leq 2K(f, \overline{U}, z_1)\log(1 + \varepsilon) + 2\log \mu_{\text{norm}}(\overline{U}, z_1) + \log(cnD^2(\log N + D + n^2)). \end{aligned}$$

Using Lemma 5.2 and (37), we finally obtain

$$\begin{aligned}
& \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} \left| \log u_*(f, \overline{U}, z_1) \right| \\
& \leq 2 \mathbb{E}_{f \in S(\mathcal{H}_{(d)})} \left(K(f, \overline{U}, z_1) \log(1 + \varepsilon) + \log \mu_{\text{norm}}(\overline{U}, z_1) \right. \\
& \quad \left. + \log(cnD^2(\log N + D + n^2)) \right) \\
& \leq \log(1 + \varepsilon) \cdot 1334978D^3N(n+1)^{D+1} \\
& \quad + D \log(\sqrt{2}(n+1)) + 2 \log(cnD^2(\log N + D + n^2)) \\
& = \mathcal{O}(D^3N(n+1)^{D+1}).
\end{aligned}$$

We observe that the initial precision $u = \frac{k_2}{nD^2(\log N + D + n^2)2(n+1)^D}$ also satisfies this inequality.

6 Proof of Theorem A

We assume now a fixed round-off unit \overline{u} . In this context we consider the following trivial variation of ALHVar:

```

Algorithm ALHFix
input  $(f, g, x)$ 
  if  $d_{\mathbb{S}}(f, g) \geq \frac{\pi}{2}$  then  $g := -g$ 
   $\tau := 0, q_{\tau} := g$ 
  repeat
    if  $\overline{u} > \frac{3}{4}\mathbf{B}(q_{\tau}, x)$  RETURN ‘‘Failure’’
     $\Delta\tau := \frac{\lambda}{d_{\mathbb{S}}(f, g)D^{3/2}\mu_{\text{norm}}^2(q_{\tau}, x)}$ 
     $\tau := \min\{1, \tau + \Delta\tau\}$ 
     $q_{\tau} := t(\tau)f + (1 - t(\tau))g$ 
     $q_{\tau} := \frac{q_{\tau}}{\|q_{\tau}\|}$ 
     $x := N_{q_{\tau}}(x)$ 
     $x := \frac{x}{\|x\|}$ 
  until  $\tau = 1$ 
RETURN  $x$ 

```

The idea is simple: the precision in ALHFix remains constant and the algorithm proceeds until either it halts returning an approximate zero x of f or it halts returning the message ‘‘Failure’’. The latter indicates that the precision is not sufficient to guarantee the correct execution of the algorithm. In the former case, we say that ALHFix *successfully halts*.

We can make ALHFix and ALHVar even closer by taking advantage of the level of generality we used to define the rounding functions r_u and the input-reading functions **read_input**. Recall, the main property of r_u is that $r_u(x) = x(1 + \delta)$ with $|\delta| \leq u$. Similarly, we have made no assumptions on the functions **read_input** _{f} besides the fact that they return a rounded-off reading of the input system f with the current precision.

For the rest of this section we assume that the rounding maps $\{r_u \mid u \in (0, 1)\}$ satisfy $r_u = r_{\bar{u}}$ for all $u \geq \bar{u}$. We also assume that, for all $h \in \mathcal{H}_{(\mathbf{d})}$ the black-box `read_inputh` is given by `read_inputh`(\cdot) = $r_u(h)$ where u is the current precision. All the results shown in Sections 4 and 5 hold in general and, a fortiori, under these assumptions as well. In addition, we have the following trivial lemma.

Lemma 6.1. *Let $h \in \mathcal{H}_{(\mathbf{d})}$ and $f = r_{\bar{u}}(h)$. If the computation of ALHFix on input $(f, \bar{U}, \mathbf{z}_1)$ successfully halts then this computation coincides with the computation of ALHVar on input $(h, \bar{U}, \mathbf{z}_1)$. In particular, they perform the same number of iterations. Otherwise, both computations coincide until a precision finer than \bar{u} is required, at which moment ALHVar proceeds but ALHFix halts with a failure message.* \square

PROOF OF THEOREM A. Algorithm MDFix is what one would expect:

Algorithm MDFix
input $f \in \mathcal{H}_{\bar{u}}$
 run ALHFix on input $(f, \bar{U}, \mathbf{z}_1)$

Because of Lemma 6.1, for $f \in \mathcal{H}_{\bar{u}}$ we have

MDFix returns ‘‘Failure’’ with input $f \iff$ for all $h \in r_{\bar{u}}^{-1}(f)$, $u_*(h, \bar{U}, \mathbf{z}_1) < \bar{u}$

where, we recall $u_*(h, \bar{U}, \mathbf{z}_1)$ is the finest u_* required by MDVar with input h . Because of the equalities $\nu_{\bar{u}}(\{f\}) = \mu(r_{\bar{u}}^{-1}(f))$ we therefore have

$$\text{Prob}_{\nu_{\bar{u}}} \{\text{MDFix returns ‘‘Failure’’}\} = \text{Prob}_{h \sim N(0, \text{Id})} \{u_*(h, \bar{U}, \mathbf{z}_1) < \bar{u}\}.$$

Theorem 4.3(iii) guarantees that, for some constant A ,

$$u_*(h, \bar{U}, \mathbf{z}_1) \geq \frac{1}{AnD^2(\log N + D + n)\mu_*^2(f, \bar{U}, \mathbf{z}_1)}.$$

Therefore,

$$\begin{aligned} & \text{Prob}_{\nu_{\bar{u}}} \{\text{MDFix returns ‘‘Failure’’}\} \\ & \leq \text{Prob}_{h \sim N(0, \text{Id})} \left\{ \mu_*^2(h, \bar{U}, \mathbf{z}_1) > \frac{1}{A\bar{u}nD^2(\log N + D + n)} \right\} \\ & \leq \text{Prob}_{h \sim N(0, \text{Id})} \left\{ K(h, \bar{U}, \mathbf{z}_1) + \frac{\log(AnD^2(\log N + D + n)\mu_{\text{norm}}^2(\bar{U}, \mathbf{z}_1))}{2\log(1 + \varepsilon)} > \frac{\log(1/\bar{u})}{2\log(1 + \varepsilon)} \right\} \end{aligned}$$

the latter by Proposition 5.3. Let

$$X := K(h, \bar{U}, \mathbf{z}_1) + \frac{\log(AnD^2(\log N + D + n)\mu_{\text{norm}}^2(\bar{U}, \mathbf{z}_1))}{2\log(1 + \varepsilon)}.$$

Then X is a positive random variable and

$$\mathbb{E}_{h \sim N(0, \text{Id})} X = \mathcal{O}(D^3 N(n+1)^{D+1})$$

by (37) and the bound $\mu_{\text{norm}}^2(\overline{U}, \mathbf{z}_1) \leq 2(n+1)^D$ (Lemma 5.2). We can now apply Markov's inequality to X —i.e., $\text{Prob}\{X > t\} \leq \frac{\mathbb{E}X}{t}$ — and we finally obtain

$$\text{Prob}_{\nu_{\overline{u}}} \{\text{MDFix returns ‘‘Failure’’}\} = \mathcal{O}\left(\frac{D^3 N(n+1)^{D+1}}{\log(1/\overline{u})}\right).$$

This shows the first assertion. To see the second, let $K_{\text{Fix}}(f, \overline{U}, \mathbf{z}_1)$ denote the number of iterations performed by ALHFix with input $(f, \overline{U}, \mathbf{z}_1)$.

Let τ_H be the value of τ when MDFix halts on input $(f, \overline{U}, \mathbf{z}_1)$, and let

$$\mu_{\bullet}(f, \overline{U}, \mathbf{z}_1) := \max_{\tau \in [0, \tau_H]} (\mu_{\text{norm}}(q_{\tau}, \zeta_{\tau})).$$

By the proof of Theorem 4.3(ii) and Lemma 6.1, the number of iterations $K_{\text{Fix}}(f, \overline{U}, \mathbf{z}_1)$ of MDFix is bounded as

$$K_{\text{Fix}}(f, \overline{U}, \mathbf{z}_1) = \mathcal{O}\left(D^{3/2} \int_0^{\tau_H} \mu_{\text{norm}}^2(q_{\tau}, \zeta_{\tau}) d\tau\right) = \mathcal{O}\left(D^{3/2} \mu_{\bullet}^2(f, \overline{U}, \mathbf{z}_1)\right). \quad (38)$$

Let j be such that the maximum $\mu_{\bullet}(f, \overline{U}, \mathbf{z}_1)$ is attained in the interval $[\tau_j, \tau_{j+1}]$. From (21),

$$\mu_{\bullet}^2(f, \overline{U}, \mathbf{z}_1) \leq (1 + \varepsilon)^2 \mu_{\text{norm}}^2(q_j, \zeta_j). \quad (39)$$

By Proposition 4.5(u) and (20), the precision \overline{u} satisfies

$$\overline{u} \leq \mathbf{B}(q, x) \leq (1 + \varepsilon)^2 \mathbf{B}(q_j, \zeta_j) = \frac{(1 + \varepsilon)^2 k_2}{nD^2(\log N + D + n^2) \mu_{\text{norm}}^2(q_j, \zeta_j)}.$$

This inequality, together with (39), imply

$$\mu_{\bullet}^2(f, \overline{U}, \mathbf{z}_1) \leq \frac{(1 + \varepsilon)^4 k_2}{nD^2(\log N + D + n^2) \overline{u}}.$$

and replacing this bound in (38) finally yields

$$K_{\text{Fix}}(f, \overline{U}, \mathbf{z}_1) = \mathcal{O}\left(\frac{1}{\sqrt{D}n(\log N + D + n^2) \overline{u}}\right). \quad \square$$

References

- [1] W. Baur and V. Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22(3):317–330, 1983.
- [2] C. Beltrán and L. M. Pardo. On Smale's 17th problem: a probabilistic positive solution. *Foundations of Computational Mathematics*, 8(1):1–43, 2008.
- [3] C. Beltrán and L. M. Pardo. Smale's 17th problem: average polynomial time to compute affine and projective solutions. *Journal of American Mathematics Society*, 22(2):363–385, 2009.
- [4] C. Beltrán and L. M. Pardo. Fast linear homotopy to find approximate zeros of polynomial systems. *Foundations of Computational Mathematics*, 11(1):95–129, 2011.
- [5] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag, 1998.

- [6] P. Bürgisser and F. Cucker. On a problem posed by Steve Smale. *Annals of Mathematics*. To appear.
- [7] F. Cucker, T. Krick, G. Malajovich, and M. Wschebor. A numerical algorithm for zero counting, I: Complexity and accuracy. *Journal of Complexity*, 24:582–605, 2008.
- [8] F. Cucker and S. Smale. Complexity estimates depending on condition and round-off error. *Journal of the American Mathematical Society*, 46:113–184, 1999.
- [9] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.
- [10] N. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 96.
- [11] M. Shub. Some remarks on Bézout’s theorem and complexity theory. In New York Springer, editor, *From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990)*, pages 443–455, 1993.
- [12] M. Shub. Complexity of Bézout’s theorem VI: Geodesics in the condition (number) metric. *Foundations of Computational Mathematics*, 9(2):171–178, 2009.
- [13] M. Shub and S. Smale. Complexity of Bézout’s theorem I: Geometric aspects. *Journal of the American Mathematical Society*, 6(2):459–501, 1993.
- [14] M. Shub and S. Smale. Complexity of Bézout’s theorem II: Volumes and probabilities. In F. Eyssette and A. Galligo, editors, *Computational Algebraic Geometry*, 109:265–285. Birkhäuser, 1993.
- [15] M. Shub and S. Smale. Complexity of Bézout’s theorem III: Condition number and packing. *Journal of Complexity*, 9:4–14, 1993.
- [16] M. Shub and S. Smale. Complexity of Bézout’s theorem V: Polynomial time. *Theoretical Computer Science*, 133:141–164, 1994.
- [17] M. Shub and S. Smale. Complexity of Bézout’s theorem IV: Probability of success; extensions. *SIAM Journal of Numerical Analysis*, 33:128–148, 1996.
- [18] S. Smale. Newton’s method estimates from data at one point. In K. Gross, R. Ewing and C. Martin, editors, *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics*, pages 265–285. Springer-Verlag, 1986.
- [19] S. Smale. Mathematical problems for the next century. In *Mathematics: frontiers and perspectives*, pages 271–294. Amer. Math. Soc., Providence, RI, 2000.